ERIC SANJUAN

Turing tests in Natural Language Processing and Information Retrieval

Contributions to the theory, development and evaluation of textual information processing systems for decision support

HDR présentée
à Avignon Université
pour obtenir le diplôme d'Habilitation à Diriger des Recherches

Laboratoire d'Informatique d'Avignon
AVIGNON UNIVERSITÉ

2018

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Computers are syntactically semantic-free machines. As Turing had anticipated, computers excel in enclosed worlds that can be described based on rules, it was expected that computers would become excellent chess players.

However Turing had also predicted that the computer would have much more difficulty with human activities in strong interaction with the external environment. This includes language and dialogue. Language learning is not based on rules but on the interaction of the learner with their emotional environment. How could a machine give the illusion of expertise in this area? And yet machines are beginning to do so through learning methods with very large data as evidenced by the main search engines on the market. It is thus possible to use these engines to ask any type of question. The first answer returned, is the most frequent, learned from a very large number of cases and based on a great volume of previous searches. It is therefore a question of mimicry rather than real intelligence.

We propose here a set of Turing tests where the computer gives the illusion that it understands the semantic content of a set of texts and responds to a need for information of a human without requiring neither large databases nor long learning processes. The tests we propose can be successful with a few well chosen rules and some heuristics. We hope to show the type of intelligence that human experts in their field can individually communicate to a machine. This is done by favouring approaches where the machine can provide a comprehensive explanation of its inference.

We propose six tests. Four in Natural Language Processing (NLP): semantic clustering SanJuan and Ibekwe-Sanjuan (2006), knowledge domain mapping SanJuan (2011), discourse segmentation da Cunha et al. (2012) and sentence specialty level recognition

da Cunha et al. (2011). Two are Information Retrieval (IR) tasks: microblog contextualization Bellot et al. (2012) and complex question answering SanJuan et al. (2010). As required by Turing, each experiment is:

- Evaluated via an interaction with a human where the computer tries to delude,
- Measurable and reproducible,
- Conducted in an open world with no explicit description.

To these constraints we add that of immediacy, the computer must answer in real time by intuition, without using large resources. In short, all algorithms must be able to be carried out in embedded mode in a small unit. The aim of this research is to help break our growing dependence on the private resources of the major search engines of the Web and the main social networks. We show that the computer can pass these six tests with these constraints.

This memoir is comprised of 2 parts.

The first part describes the four tests we propose in NLP over three chapters (Chapter 2 to Chapter 4).

The second chapter on semantic classification without knowledge is based on the TermWatch system developed with Fidelia Ibekwe - SanJuan and with the support of INIST - CNRS from 2002 to 2006.

The next chapter on knowledge domain mapping of a domain is based on the latest algorithms added to this system. These are based on the thesis of Bangaly Kaba that we co-supervised with Anne Berry of LIMOS - Clermont Ferrand 2. This collaboration with Anne Berry was extended to Alain Sigayret to explore the relationship between graph algorithms and computation of association rules also integrated in TermWatch.

The fourth chapter brings together the following two tests on discourse segmentation and the recognition of a specialty level of an isolated sentence. This is a joint effort with Iria da Cuhna of the Pompeu Fabra University (UPF) in Barcelona and Juan Manuel Torres Moreno in collaboration with Professors Maria Theresa Cabré (UPF) and Gerardo Sierra Autonoma University of Mexico (UNAM). This research continued with the PHD of Alejandro Molina on the compression of sentences from a discrete segmentation.

The second part is devoted to information retrieval (IR), or rather to the characterization of IR tasks that can benefit from NLP. The classic models and tests of IR have indeed moved away from NLP.

Chapter 5 takes up the classical task of interactive query reformulation and shows the pivotal role of terminology extraction. The machine proposes terms and is responsible for inserting them in the initial query. This minimal interaction is sufficient to significantly improve the results without requiring large resources or historical data. It is an obvious alternative to saving personal data.

The last chapter is devoted to the QA and Contextualization tasks of INEX and CLEF evaluation campaigns. We proposed and defined these tasks within the framework of the CAAS project funded by the ANR and led by Professor Josiane Mothe of IRIT with Patrice Bellot for the LIA. Animation and follow-up of these tasks was done in collaboration with Véronique Moriceau and Xavier Tannier of LIMOS. The evaluation metrics initially used are those presented and studied in Saggion et al. (2010). In this chapter we review the definition and main results of the 2009-2011 editions as well as the basic system that was proposed to the participants. This system is inspired by the ideas and results of the thesis of Sylvia Fernandez that we co-supervised with Juan Manuel Torres Moreno and Romain Deveaud that we co-supervise with Patrice Bellot.

Part I

Natural Language Processing

Chapter 2

Semantic Classification without Learning

2.1 Introduction

We developed a fast and efficient text mining system that builds clusters of noun phrases (multi-word terms) without need of document co-occurrence information. This is useful for mapping out research topics at the micro-level. Because we do not consider the within document co-occurrence, our approach can be conceived as an *out-of-context clustering* except if we consider the *intra-term* context, i.e., words appearing in the same terms can be said to share a similar context. Terms are clustered depending on the presence and number of shared linguistic relations. For instance, a link will be established between the two terms *humoral immune response* and *humoral Bhx immune response* since one is lexically included in the other. Likewise *clustering algorithm* is linked to *computer algorithm* by a modifier substitution. This lexico-syntactic approach is suitable for clustering multi-word text units which rarely re-occur *as is* in the texts. Such multi-word terms (MWTs) often result in very large and sparse matrices or graphs[1] that are difficult to handle by the existing approaches to clustering which rely on high frequency information. The resulting system, called TermWatch (Ibekwe-SanJuan, 1998a; Sanjuan et al., 2005) can be applied to several tasks like domain topic mapping, text mining, query refinement or question-answering (Q-A).

Some attempts have been made to cluster document contents in the bibliometrics,

[1] In the experiments run up to date, we have been able to handle graphs of 80,000 terms in real time applications for online data analysis and query refinement.

scientometrics and informetrics fields. Some authors have considered the clustering of keywords, classification codes or subject headings assigned to documents by indexers (Callon et al., 1991; Zitt and Bassecoulard, 1994; Braam et al., 1991). Although these information units depict the thematic contents of documents, they are external to the documents themselves and do not allow for a fine-grained analysis of the current topics addressed in the full texts. In studies where the document contents were considered, only lone words were extracted through statistical analysis. The majority of clustering methods used in the information retrieval field (Eisen et al., 1998; Cutting et al., 1992; Karypis et al., 1994) are also based on the vector-space representation model of documents (bag-of-words approach). To reduce the dimensions of the vector space, words with a discriminating power are selected based on term weighting indices like the *Inverse Document Frequency* (IDF), *Mutual Information* (MI) or the cosinus measure. This also results in the drastic elimination of more than half of the initial data from the analysis. Our text mining approach treats highly frequent and low frequent terms equally. This is important for applications like science and technology watch where the focus is on novel information often characterised by low frequency units (weak signals). Price and Thelwall (2005) have demonstrated the usefulness of low frequency words for scientific web intelligence (SWI). They showed that removing low frequency words reduced cluster coherence and separation, i.e., clusters were less dissimilar.

Glenisson et al. (2005) proposed combining full text analysis with bibliometric analysis in order to cluster the research themes of 85 scientific papers. Text contents were represented as vectors of lone words. Stemming was performed on the words and bigrams were detected, i.e. sequences of two adjacent words that occurred frequently. It is a well known fact that stemming brutally removes the semantics of derived or inflected words. For instance, "*stationary, station, stationed*" are all reduced to *station*. Also, bigrams may not always correspond to valid domain terms. The authors weighted the bigrams using the Dunning likelihood ratio test (Dunning, 1993). This led to selecting the 500 topmost bigrams for analysis and discarding the rest. One of the interesting findings of this study is that clustering items from full texts rather than keywords or terms from the reference section leads to a more fine-grained and accurate mapping of research topics. This finding is in line with our text mining approach.

Polanco et al. (1995) developed the Stanalyst informetrics platform. Stanalyst comprises a linguistic component which identifies variants of MWTs used to augment their occurrences. The MWTs are then clustered based on document co-occurrence information. To the best of our knowledge, no informetric method has considered clustering phrases based on linguistic relations. The TermWatch approach is based on the hypothesis that clustering multi-word terms (MWTs) through lexico-syntactic and semantic relations can yield meaningful clusters for various applications. In view of this, we

developed a methodology that can handle very large and sparse matrices in real time. For instance, in the current experiment, the input list of terms is 31,398, none which is eliminated prior to the matrix reduction phase.

The clustering algorithm implemented in TermWatch is named CPCL (Classification by Preferential Clustered Link). This algorithm was first published in (Ibekwe-SanJuan, 1998a) but owing to its fundamental differences with existing approaches, setting up an adequate comparison framework with other methods has been a bottleneck issue. In this chapter, we focus on the evaluation with other clustering algorithms (variants of partitioning and hierarchical algorithms). Evaluation is carried out on a test corpus (the GENIA project) which comes with an answer key (gold standard). This will ensure that the results being presented are grounded in the real world.

The rest of the chapter is organized as follows: section 2.2 gives details of the test corpus; section 2.3 describes our text mining methodology; section 2.4 presents the evaluation method; section 2.5 describes the experimental setup; section 2.6 discusses the results of the evaluation with other clustering methods; section 2.7 draws remarks and conclusions.

2.2 Test corpus

In order to carry out an evaluation, we chose a dataset with an existing *ideal partition* (gold standard). The GENIA project[2] consists of 2,000 abstracts downloaded from the MEDLINE database using the search keywords: *Human, Blood Cells,* and *Transcription Factors*. Biologists manually annotated the valid domain terms in these texts, yielding 31,398 terms. This ensures in our experiment that competing methods start from the same input. The GENIA project also furnished a hand-built ontology, i.e. a hierarchy of these domain terms arranged into semantic categories. There are 36 such categories at the leaf nodes. Each term in the GENIA corpus was assigned a semantic category at the leaf node of the ontology. We shall refer to the leaf node categories as *classes* henceforth. Of course, the GENIA ontology's hierarchy, the number of classes and the semantic category of each term were hidden from the clustering methods. It should be noted that since the GENIA ontology is a result of a human semantic and pragmatic analysis, we do not expect automatic clustering methods to reproduce it exactly without prior and adequate semantic knowledge. The goal of the evaluation is to determine the method whose output requires the least effort to reproduce the classes at the leaf nodes of the ontology. Also, it is worth noting that although the authors of this project use

[2] *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.*

the term *ontology* to qualify this hierarchy, it is more of a small taxonomy. Indeed, the GENIA *ontology* is still embryonic because of its small size (36 classes, 31,398 terms). The classes are of varying sizes. The largest class, called *other name* has 10,505 terms followed by the *protein molecule* class with 3,899 terms and the *dna domain or region* class with 3,677 terms. The 12 smallest classes (*rna domain or region inorganic, rna substructure, nucleotide, atom, dna substructure, mono cell, rna n/a, protein n/a, carbohydrate, dna n/a, protein substructure*) each has less than 100 terms. It is quite revealing that the largest class is a miscellaneous class. This suggests that this class can be further refined. Also some relations normally found in a full-fledged ontology are absent (synonymy in particular). This tends to suggest that this hierarchy is a weaker semantic structure than an ontology and can thus constitute an adequate clustering task. For these reasons, we prefer to refer to it as the *GENIA taxonomy* henceforth.

Table 2.1 gives some examples of terms in the GENIA corpus.

GENIA Category	Terms
amino acid monomer	amide-containing amino acid asparagine n-acetylcysteine
atom	cytosolic calcium feca2+
body part	organ peripheral lymphoid organ tumor-draining lymph node
cell component	1389 sites/cell b6d2f1 mouse uterine cytosol cytoplasmic protein extract il-13-treated human peripheral monocyte nuclear extract
cell line	anergized t cell adherence-isolated monocyte xenopus hepatocyte
other name	anatomic tumor size apoptosis follicular lymphoma

Table 2.1: Examples of terms in GENIA corpus

Figure 2.1 shows the fast decreasing distribution of terms in the 35 classes. We omitted the largest class, called *other name* which concentrated 33% of the terms because it was difficult to fit in. A few number of classes (*protein molecule, dna domain or region, protein family or group, cell line, cell type*) concentrated the rest of the terms (almost 75%). The bars show the proportion of terms according to their length. As a consequence of this fast decreasing model, a clustering method optimised for one of the prominent classes can obtain good scores without correctly classifying terms in the

majority of the smaller classes. Another feature that can be observed in figure 2.1 is that the distribution of one word terms is not correlated with the general distribution of terms. Meanwhile, we will see in section §2.6 that most of the clustering methods perform better on long terms and thus on classes like "*protein family or group*" and "*dna domain or region*" that contain few one word terms. In an OTC task, the intrinsic properties of MWTs (like term length) obviously play an important role since they are the only available context.

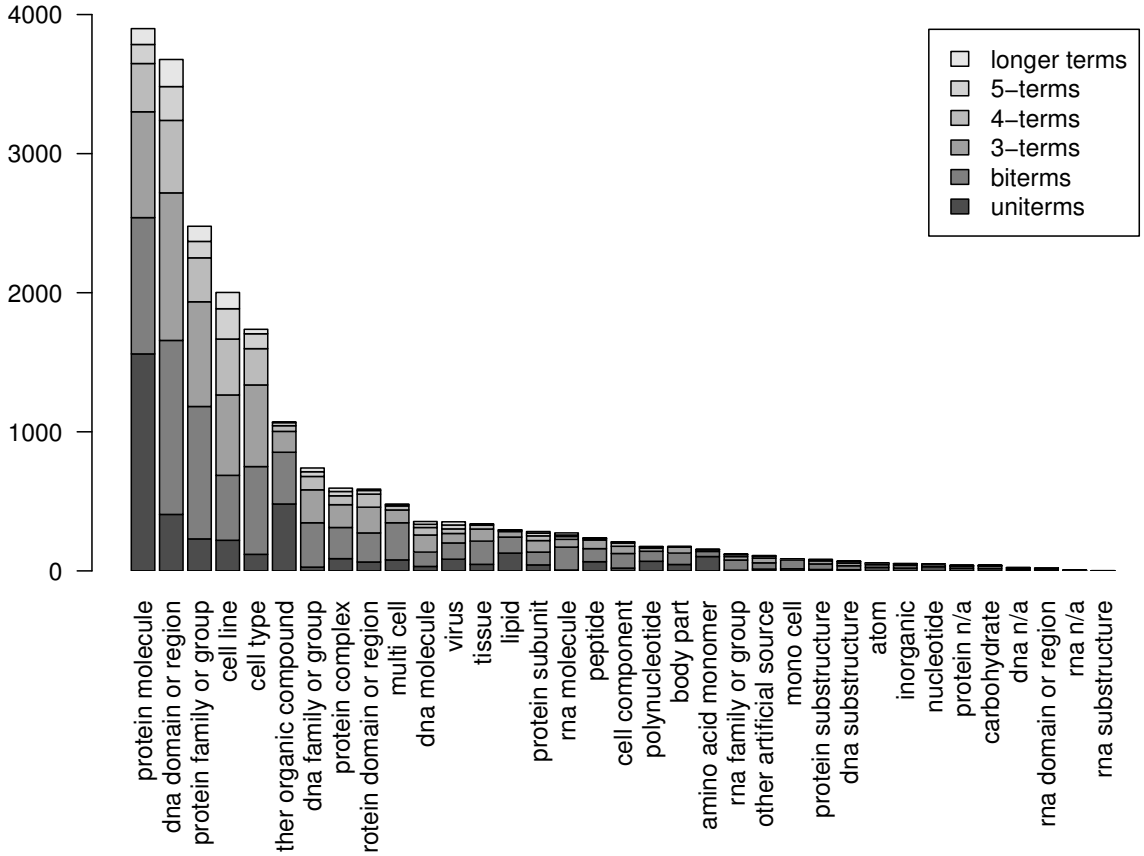

Figure 2.1: Distribution of terms in GENIA categories.

2.3 Overview of our text mining methodology

Our methodology consists of three major components: MWT extraction; relation identifier and clustering module. An integrated visualisation package[3] can be used if topic mapping is the target. In this experiment, this aspect will not be explored as evaluation will focus on cluster quality and not on their layout. However, interested readers can find an application of research topic mapping in (Ibekwe-SanJuan and SanJuan, 2004).

[3]The aiSee visualization package (http://www.aisee.com) has been integrated to the system.

2.3.1 Term extraction module

Note that in the current experiment, our term extraction module was not used as the terms were already manually annotated in the corpus. We however describe summarily its principle. TermWatch performs term extraction based on shallow natural language processing (NLP) techniques. Extraction is implemented via the NLP package developed by the University of Edinburgh. LTPOS is a probabilistic part-of-speech tagger based on Hidden Markov Models. It uses the Penn Treebank tag set which ensures the portability of the tagged texts with many other systems. LTCHUNK identifies simplex noun phrases (NPs), i.e., NPs without prepositional attachments. In order to extract more complex terms, we wrote contextual rules to identify complex terminological NPs. About ten such contextual rules were sufficient to take care of the different syntactic structures in which nominal terms appear in English. Given that some domain concepts can appear as long sequences like in *parental granulogyte-macrophage colony-stimulating factor (GM-CSF)-dependent cell line*, it is obvious that such MWTs are not likely to re-occur frequently in the corpus. Hence, the difficulty of clustering them with methods based on co-occurrence criteria.

2.3.2 Relation identifier

Different linguistic operations can occur within NPs. These operations either modify the structure or the length of an existing term. They have come to be known as *variations* and have been well studied in the computational terminology field (Jacquemin, 2001; Ibekwe-SanJuan, 1998b). Variations occur at different linguistic levels: morphological (gender and spelling variants), lexical (substitution of one word by another in an existing term), syntactic (expansion or structural transformation of a term), semantic (synonyms, generic/specific relations). Our relation identifier tries to acquire all these types of variations among the input terms.

Morphological variants

These refer to number (*tumor cell nuclei* /*tumor cell nucleus*) and gender variations in a term and also to spelling variants. They enable us to recognize different appearances of the same term. For instance, *IL-9-induced cell proliferation* will be recognized as a spelling variant of *IL 9-induced cell proliferation*. Spelling variants are identified using cues such as special characters while gender and number variants are identified using WordNet (Fellbaum, 1998)

Lexical variants

We call substitution variants operations involving the change of only one word in a term, either in the modifier position (*coronary heart disease* ↔ *coronary lung disease*) or in the head position (*mutant motif* ↔ *mutant strain*). The head is the noun focus in an English NP, i.e., the subject while the modifier plays the role of a qualifier (an adjective). The head word is usually the last noun in a compound phrase (strain in *mutant strain*) or the last noun before a preposition in a prepositional structure (retrieval in *retrieval of information*).

Syntactic variants

These refer to the addition of one or more words to an existing term as in *information retrieval* and *efficient retrieval of information*. We call these operations *expansions*. Expansions that affect the modifier words are further broken down into left-expansion and insertion. Alternatively, expansions can affect the head word. In this case, we talk of *right expansion*.

Morphological variants (spelling) and permutation variants are recognized first since they refer to the same term. Then these variants are used to recognize the more complex variants. For instance, *B cell development* haven been recognized as a spelling variant of *B-cell development*, this enables the identification of other types of variants (syntactic and lexical) containing the two spelling variants. Variations are assigned a role during clustering depending on their interpretation. This will be further detailed in section §2.5.

Semantic variants

It is an accepted fact that syntactic relations suggest semantic ones (left expansions and insertion can engender *generic-specific* links, some substitution variants can reflect *see also* relations). However, these semantic relations are not explicit. Moreover, the types of relations considered so far all require one stringent condition: that the related terms share some common words. This leaves out terms which can be semantically-linked but without sharing common words, i.e. synonyms. In order to acquire explicit semantic links, we need an external semantic resource. For this purpose, we chose WordNet (Fellbaum, 1998), a large coverage semantic database which organizes English words into synsets. A synset is a particular sense of a given word. Since WordNet organizes

only words and not multi-word terms, we had to devise rules in order to map *word-word* semantic relations into *"MWT- MWT"* relations in our corpus. One way to achieve this is to replace words by their synsets and then apply the same variation relations to sequence of synsets. However, like all external resources, WordNet has some limitations. First is its incompleteness vis-à-vis specialised domain terminology. Second, being a general purpose semantic database, WordNet establishes links which can be incorrect in a specialized domain.

We thus restricted the use of WordNet to filtering out lexical substitutions, and consequently to pairs of terms that share at least one word in order to reduce the number of wrong semantic links. Only a very few number of relations were found. The following rule was applied to lexical substitutions in order to identify the semantic ones using WordNet hierarchy: given two terms related by a lexical substitution, check if the two words substituted are linked by an ascending or descending path in the hierarchy. Observe that, by definition of lexical substitutions, this rule ony applies to words that are in the same grammatical position (head or modifier).

In this way, we acquired the following synonymy relations:

$$\begin{array}{rcl} \textit{T cell growth} & \sim & \textit{T cell maturation} \\ \textit{antenatal steroid treatment} & \sim & \textit{prenatal steroid treatment} \end{array}$$

Only 365 WordNet modifier substitutions and 208 WordNet head substitutions were found whereas lexico-syntactic variants were much more abundant (see table 2.2 below).

Table 2.2 gives the number of variants identified for each type among the GENIA terms. As a term can be related to many others, the number of relations is always higher than the number of terms.

Variation relation	Terms	Relations
Spelling variants	1560	2442
Left Right-expansions (exp_2)	294	441
Right-expansions (exp_r)	2329	3501
Left-expansions (exp_l)	2818	4260
Insertions (ins)	526	798
Modifier-substitutions (sub_mod.)	4291	37773
Head-substitutions (sub_head)	781	1082
WordNet-synonyms (sub_wn)	365	208

Table 2.2: Statistics on variation relations per type

2.3.3 Clustering module

The TermWatch system implements a graph-based approach of the hierarchical clustering called CPCL (Classification by Preferential Clustered Link) originally introduced by Ibekwe-SanJuan (1998b). The main features of this approach are :

1. the intuitiveness of its results for human users since any pair of terms clustered together are related by a relative short path of real linguistic relations,
2. an ultrametric model that ensures the existence of a unique and robust solution,
3. its linear time complexity on the number of variations that allows interactive data analysis since clustering can be processed in real time.

We show here that this algorithm can be applied to other types of inputs. For that, we need to cast the description of the algorithm in the more general context of data analysis.

Let S be a sparse similarity data matrix defined on a set Ω of objects. This matrix can be represented advantageously by a valuated graph $G = (\Omega, E, s)$ where E is the set of *edges* made of all unordered pairs $\{i, j\}$ of objects such that $S_{ij} > 0$ and s is the valuation of edges defined for all $(i, j) \in \Omega^2$ by $s(i, j) = S_{ij}$. In the case of sparse data, the size of E is much smaller than $|\Omega|^2$.

Let Val_S be the set of values in S. If $|Val_S| \ll |\Omega|$ then, the usual hierarchical algorithms will produce small dendrograms since they will have at most $|Val_S|$ levels. Thus, they will induce a very reduced number of intermediary balanced partitions in the gap between the trivial discrete partition and the family of connected components of G. A way to correct this drawback of hierarchical clustering without losing its intuitiveness and computer tractability is not to consider edge values in an absolute way but in the context of adjacent edges. Thus, two objects related by an edge e will be clustered at a given iteration, only if the value of e is greater than any other value in its neighborhood. This means that i, j will be clustered at the first iteration only if S_{ij} is greater than the maximum in the line $S_{i.}$ and in the column $S_{.j}$. It has been shown in Berry et al. (2004) that this variant of hierarchical clustering preserves its main ultrametric properties.

This solution is specially well adapted when the observed similarities between objects are generated by pairwise observations. In the case of out-of-context clustering (OTC), given three terms u, v, t such that v shares at least one word with u and t (possibly not the same), we will consider a local criteria to decide if v is closest to u or to t.

In this approach, the clustering phase can be easily implemented using the following straightforward procedure which we call $SLME$ (Select Local Maximum Edge). This procedure runs in linear time on the number of edges. In fact, the procedure does as many comparisons as the sum of vertex degrees which is two times the number of vertices. It uses a hash table m to store, for each vertex x, the maximal value of previously visited adjacent edges.

```
SLME procedure

Input  : a valued graph (V,E,s)
Output : a relation R on V
L:={}
D:={}

for every x in V, m[x] := -1

while V-L is not empty
  Select one vertex v in V-L
  add v to L
  C:={v}
  while C is non empty
    x:=pop(C)
    add x to L
    add neighbours(x) - L to C
    m[x] := max{s(n): n in neighbours(x)}
    for every n in neighbours(x)
    if m[n]=m[x] add {n,x} to R
```

Once done, the clustering phase consists in computing the reduced graph $G/_R$, whose vertices are the connected components of the subgraph (V, R) of G and in inducing a new valuation according to a hierarchical criteria chosen among the following:

single-link: the value of an edge in $G/_R$ between two components C_1, C_2 is the maximal value of edges in $E_{C_1,C_2} = E \cap (C_1 \times C_2)$.

complete-edge: the minimal value in $E_{C_1C_2}$,

average-edge: the average value in $E_{C_1C_2}$,

vertex-weight: the sum of values in $E_{C_1C_2}$ over $|C_1| + |C_2|$

Observe that the above *complete-edge* and *average-edge* criteria differ from the usual complete and average link clustering since they are computed on a restricted set of pairs. The *vertex-weight* criterion is the one that best minimized the chain effect in our experiments. However in general, single link will also be satisfactory because the chain effect has already been reduced by the SLME procedure. In fact, this approach appears to be robust with regard to the clustering criteria. It is more sensitive to the existence of very small values in the similarity matrix S. Indeed, any non null value will generate an edge in the graph and if this edge is the only one linking two objects, then these objects will be clustered even if the similarity is very small. This drawback can be corrected by the use of a threshold which clarifies the borderline between null values and significant similarities.

The CPCL algorithm then becomes:

```
Algorithm CPCL

input      : a valued graph G=(V,E,s)
parameters : a threshold t and a number of iterations I
output     : a partition of V
for i=1 to i=I do
  E':= {e in E : v(e) > t}
  R := SLME(V,E',s)
  G := G/R
return V
```

It involves I calls to the SLME procedure on the current reduced valued graph (V,E',s).

It follows from this re-exploration of CPCL that it can by used for fast clustering of sparse similarity matrix with a reduced range of distinct values.

Until now, this algorithm has been applied to the following similarity matrix defined on groups of objects and generated in two steps:

Step1: we consider a reduced subset of variation relations among those presented in subsection 2.3.2 that we shall note $COMP$.

We then compute the set of connected components generated by the $COMP$ relations. Terms that are not related by any of the variations in $COMP$ will form singleton components.

Step2: We select a second subset of variations denoted by $CLAS$ to group components. Next, given two components C_1 and C_2, a similarity value v is defined in the following way:

$$v = \sum_{R \in CLAS} \frac{|R \cap C_1 \times C2|}{|R|}$$

This similarity relies on the number of variations across the components and on the frequency of the variation type which on a large corpus will substantially reduce the influence of the most noisy variations like lexical substitutions on binary terms. The resulting matrix has all the characteristics that justify the application of the CPCL algorithm.

2.3.4 Implementation issues

Figure 2.2 gives an overall view of the system. It is currently run on-line on a Linux Apache MySQL Php PERL Secured (LAMPS) server[4]. The three components term extractor, relation identifier and clustering module are implemented as PERL5 OO programs while all the data are stored in a MySQL database. Clustering outputs can be accessed either via an integrated visualization package (aiSee based on Graph Description Language) for domain topic mapping or through an interactive hypertext interface based on PERL DBI and CGI packages. This interactive interface enables the user to browse the results, from the term network(variation links) to clusters contents and finally to documents where the terms appeared. The systems' modules can also be executed from this interface.

2.4 Evaluation metrics

Evaluating the results of a clustering algorithm remains a bottleneck issue (Yeung and Ruzzo, 2001; Jain and Moreau, 1987; Tibshirani et al., 2000). The objective of the evaluation for our specific task : clustering multi-word terms out-of-context, is detailed in §2.4.1 followed by a review of existing evaluation methods §2.4.2. Finally, in §2.4.3 we propose enhancements to the editing distance suggested by Pantel and Lin (2002) for cluster evaluation.

[4]TermWatch is available for research purposes after obtaining an account and a password from the authors.

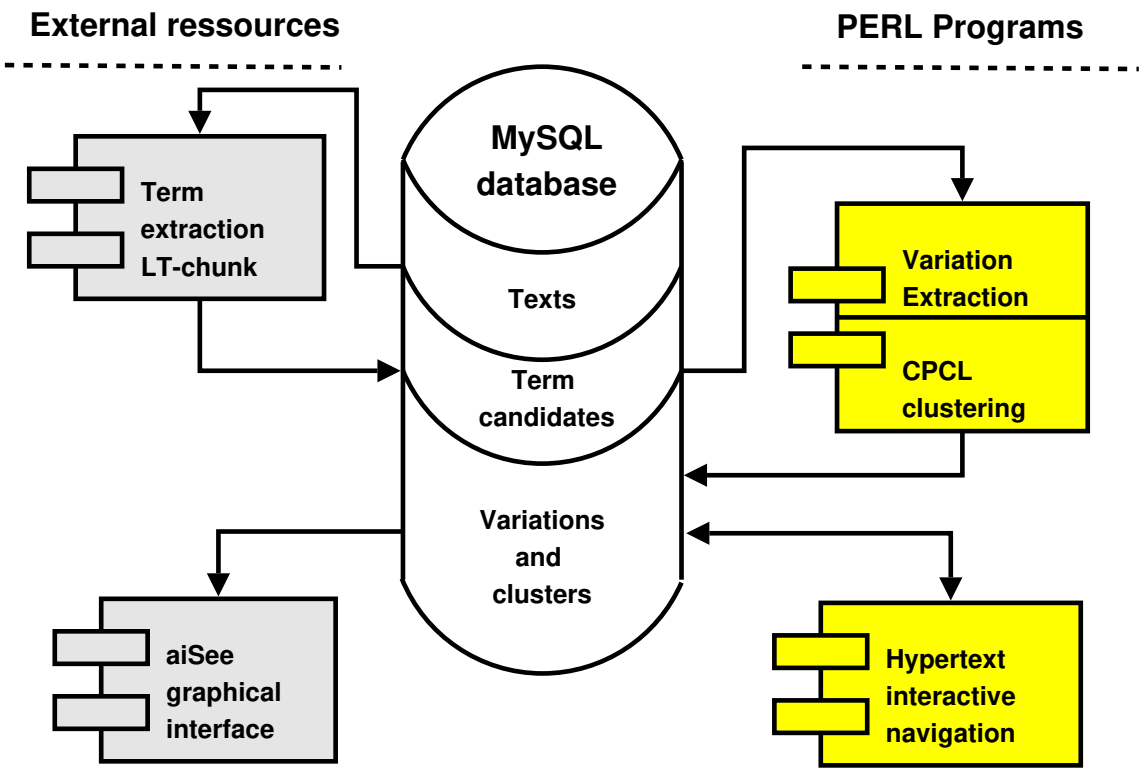

Figure 2.2: Overall view of the TermWatch system

2.4.1 Out-of-context Term Clustering (OTC)

Given a list of terms, the task consists in clustering them using exclusively surface lexical information in order to obtain coherent clusters. In this framework, clustering is done without contextual document information, without any training set and in a completely unsupervised way. We refer to this task as OTC (Out-of-context Term Clustering).

Let us emphasize that OTC is different from Entity Name Recognition (ENR). ENR task as described in Kim et al. (2004) is based on massive learning techniques and new terms are forced to enter known categories. Whereas in unsupervised clustering, a new cluster can be formed of terms not belonging to an already existing category. This can lead to the discovery of new domain topics. It should also be noted that MWTs cannot be reduced to single words. Unlike single words, a MWT can occur only once "as is" (without variations) in the whole corpus. It is thus difficult for the usual *document* × *feature* representation to find enough frequency information to form clusters. Therefore methods based on *term-document* representation cannot be directly applied to OTC without adaptation. This adaptation is described in further details in §2.5.

2.4.2 Existing measures for cluster evaluation

Cluster evaluation generally falls under one of these two frameworks:

1. intrinsic evaluation: evaluation of the quality of the partitions vis-à-vis some criteria.
2. extrinsic evaluation : task-embedded evaluation or evaluation against a gold standard.

Intrinsic evaluation, also called "internal criteria" is used to measure the intrinsic quality of the clusters in the absence of an external ideal partition. Internal criteria concern measures like cluster homogeneity and separation, or the stability of the partitions with respect to sub-sampling (Hur et al., 2002). Alternatively, the measure can also seek to determine the optimal number of clusters (Hur et al., 2002).

Extrinsic evaluation, also known as "external criteria" refers to the comparison of a partition against an external ideal solution (gold standard) (Milligan and Cooper, 1985; Jain and Moreau, 1987) or a task-embedded evaluation. The comparison with a gold standard is done using measures like the Rand index or its adjusted variant (Hubert and Arabie, 1985) that measures the degree of agreement between two partitions[5]. Milligan and Cooper (1986) recommended the use of Adjusted Rand index even when comparing clusters at different levels of the hierarchy. As observed by Yeung and Ruzzo (2001), external criteria has the advantage of providing an *"independent unbiased assessment of the cluster"* but has as inconvenience the fact that they are hardly available.

Internal criteria has as advantage the fact that it can bypass the necessity of having an external ideal solution but its major inconvenience is that evaluation is based on the same information from which the clusters were derived. Pantel and Lin (2002) observed a flaw in the external criteria approach as suggested by the Rand index. According to them, computing the degree of agreements and disagreements between proposed partitions and an ideal one can lead to unintuitive results. For instance, if the ideal partition has 20 equally-sized clusters with 1000 elements each, treating each element as its own cluster will lead to a misleading high score of 95% . We observe also that the Rand index and the adjusted Rand Index (Hubert and Arabie, 1985) have the following flaws:

[5]Given two equivalence relations P and Q defined on a set Ω, the rand Index is the number of agreements between the two relations $|(P \cap Q) \cup \neg(P \cup R)|$ over the total number of pairs $|\Omega|^2$. The adjusted rand index assumes the generalized hypergeometric distribution as the model to ensure that two random partitions do take a constant null value.

- they are computationally expensive since they require $|\Omega|^2$ comparisons which is problematic when $|\Omega|$ is large,
- they are too sensitive to the number of clusters when comparing clustering outputs of different size (Wehrens et al., 2003),
- the adjusted Rand Index supposes a hyper-geometric model which is obviously not fitted to the distribution of terms in the current experiment (GENIA categories).

Denoeud et al. (2005) tested the ability of different measures in determining the distance between two partitions. The Jaccard measure appeared as the best in this task since it does not have the drawbacks of the (adjusted) Rand Index. It computes the number of pair of items clustered together by two algorithms divided by the total number of pairs clustered by one of the algorithms. However, it cannot take into account the specificity of a target distribution. More precisely, suppose that we want to measure the gap between a clustering output and a target classification, suppose moreover that the target classification has a very large class with a great number of terms whereas the mean size of the other classes is small, (this is precisely the case in the GENIA taxonomy where the *other name* class groups 33% of all the terms in this taxonomy), although this class is disproportioned, it is definitely not the most informative. The Jaccard measure will favour methods that focus on the detection of the biggest class against more fine-grained measures that try first to fit the distribution of items in the smaller classes. Yeung and Ruzzo (2001) proposed a compromise for cluster evaluation in which evaluation is based on the predictive capacity of the methods vis-à-vis a hidden experimental condition. They tested their method on gene expression (microarray) data. This approach, aside from being computationally intensive, is not suitable for datasets where no experimental conditions (hidden or otherwise) obtain nor will it be suitable for datasets where the different samples do not share any dependent information.

In the task-embedded evaluation framework, what is evaluated is not the quality of the entire partition but rather that of the *best cluster* (Pantel and Lin, 2002), i.e., the cluster which enables the user to best accomplish his information seeking need. This is typically the case with cluster evaluation in the information retrieval field.

Following the extrinsic evaluation approach, Pantel and Lin (2002) proposed the use of the editing distance to evaluate clustering outputs. The idea is to evaluate the *cost* of producing the ideal solution from the proposed partitions. This supposes the existence of an external ideal solution. The editing distance is an old notion used to calculate the cost of elementary actions like *copy, merge, move, delete* needed to obtain one word (or phrase or sentence) from another. Here, the authors applied it to cluster contents and

chose to consider three elementary actions: copy, merge, move. Considering the OTC task, we needed a measure that focused on cluster quality (homogeneity) vis-à-vis an existing partition (here the GENIA classes). Pantel & Lin's editing distance appeared as the most suitable for this task. It is adapted to the comparison of methods producing a great number of clusters (hundreds or thousands) and of greatly differing sizes. On a more theoretical level, the idea of *editing distance* is conceptually suited to the nature of our evaluation task, i.e., calculate the *effort* or the *cost* required to attain an existing partition from the ones proposed by automatic clustering methods.

2.4.3 Metrics for evaluation of clusters

Given an existing target partition, Pantel and Lin (2002)'s measure evaluates the ability of clustering algorithms to detect part of the structure represented by this partition. This measure extends the notion of editing distance to general families of subsets of items. In particular, it allows to consider fuzzy clustering where clusters overlap (copy action). Here we will not use this feature since we target crisp clustering. Hence, we focus on the two elementary operations : *merges* which is the union of disjoint sets and *moves* that apply to singular elements. In this restricted context, Pantel and Lin (2002)'s measure has a more deterministic behaviour and shows some inherent bias which we will correct.

To measure the distance between a clustering output and an ideal partition, these authors considered the minimal number of merges and moves that have to be applied to a clustering output in order to obtain the target partition. In fact, this number can be easily computed since the number of merges corresponds to the number of extra-classes and the number of moves to the number of elements that are not in the dominant class of the cluster. Indeed, each cluster is associated to the class with which it has the maximum intersection. The elements of a cluster which are not in the intersection will then have to be moved.

Thus, let Ω be a set of objects for which we know a crisp classification $\mathcal{C} \subseteq 2^{\Omega}$, seen as a family of subsets of Ω such that $\bigcup \mathcal{C} = \Omega$ and $C \cap C' = \emptyset$ for all C, C' in $\mathcal{C}$. Consider now a second disjoint family $\mathcal{F}$ of subsets of Ω representing the output of a clustering algorithm. For each cluster $F \in \mathcal{F}$, we denote by $\mathcal{C}_F$ the class $C \in \mathcal{C}$ such that $|C \cap F|$ is maximal. Pantel & Lin's measure can be re-formulated thus:

$$\mu_{LP}(\mathcal{C}, \mathcal{F}) \quad = \quad 1 - \frac{(|\mathcal{F}| - |\mathcal{C}|) + \sum_{F \in \mathcal{F}} (|F| - |\mathcal{C}_F \cap F|)}{|\Omega|} \tag{2.1}$$

In the numerator of formula 2.1, the term $|\mathcal{F}|-|\mathcal{C}|$ gives the number Mg of necessary merges, and the sum $\sum_{F\in\mathcal{F}}(|F|-|\mathcal{C}_F\cap F|)$ the number Mv of moves. The denominator $|\Omega|$ of (2.1) is supposed to give the maximal cost of building the classification $\mathcal{C}$ from scratch. Indeed, Pantel & Lin considered two trivial partitions: the discrete one where all clusters are singletons (every term is its own cluster) and the complete one where all terms are in a single cluster. These trivial partitions are supposed to be at equal distance from the target classification. These authors suggest that the complete clustering needs $|\Omega|$ moves and the discrete $|\Omega|$ merges but this turns out not to be the case.

Clearly, discrete clustering only needs $|\Omega|-\mathcal{C}$ merges. Moreover, if $g=\max\{|C|:C\in\mathcal{C}\}$ is the size of the largest class in $\mathcal{C}$, then the distance of the trivial complete partition to the target partition is $|\Omega|-g$. It follows that in the case where g is much more greater than the mean size of classes in $|\mathcal{C}|$, Pantel & Lin's measure, based on the total number of necessary moves and merges over $|\Omega|$ favours the trivial complete partition over the discrete one and therefore algorithms that produce very few clusters, even of poor quality. Incidentally, this happens to be the case with the GENIA classes. Following these observations, we propose the following corrected version (2.2) where the weight of each move is no more 1 but $|\Omega|/(|\Omega|-g)$ and the weight of a merge is $|\Omega|/(|\Omega|-|\mathcal{C}|)$:

$$\mu_{ED}(\mathcal{C},\mathcal{F}) = 1-\frac{\max\{0,|\mathcal{F}|-|\mathcal{C}|\}}{|\Omega|-|\mathcal{C}|}-\frac{\sum_{F\in\mathcal{F}}(|F|-|\mathcal{C}_F\cap F|)}{|\Omega|-g} \tag{2.2}$$

$$= 1-\frac{Mg}{|\Omega|-|\mathcal{C}|}-\frac{Mv}{|\Omega|-g} \tag{2.3}$$

The maximal value of μ_{ED} is 1 in the case where the clustering output corresponds exactly to the target partition. It is equal to 0 in the case that $\mathcal{F}$ is a trivial partition (discrete or complete).

However, μ_{ED} can also take negative values. Indeed consider the extreme case where $\mathcal{C}$ is of the form $\{A,B_1,...,B_n\}$ with one class $A=\{\alpha_1,...,\alpha_n,\omega_1,\omega_2\}$ with $n+2$ elements and n singleton classes $B_i=\{\beta_i\}$. Now take as $\mathcal{F}$ the whole family of n pairs $\{\alpha_i,\beta_i\}$ for $1\leq i\leq n$ augmented with the singletons $\{\omega_1\},\{\omega_2\}$. Then:

$$\mu_{ED}(\mathcal{C},\mathcal{F}) = 1-\frac{1}{(n+n+2)-(n+1)}-\frac{n}{(n+n+2)-(n+2)}$$

$$= -\frac{n}{n+1}<0$$

and $\lim_{n\to\infty}\mu_{ED}(\mathcal{C},\mathcal{F})=-1$

In fact, in the case that g is much more greater than the mean size of classes and that the distribution of sizes of classes fits an exponential model, we have experimentally checked that $\mu_{ED}(\mathcal{C}, \mathcal{F}) \in]-1, 0[$ for random clusterings $\mathcal{F}$ with $2g$ clusters and equiprobability for an element ω to be affected to anyone of these clusters.

Based on the corrected μ_{ED} index, we propose a complementary index, Cluster homogeneity (μ_H) defined as the number of *savings* (product of μ_{ED} per $|\Omega|$) over the number Mv of movings:

$$\mu_H(\mathcal{C}, \mathcal{F}) \quad = \quad \frac{\mu_{ED}}{1 + Mv} \times |\Omega|$$

μ_H takes its maximal value $|\Omega|$ if $\mathcal{F} = \mathcal{C}$ and, like the μ_{ED} measure, it is null if $\mathcal{F}$ is one of the two trivial partitions.

We will use μ_H to distinguish between algorithms having similar editing distances but not producing clusters of the same quality (homogeneity). However, since the cluster homogeneity measure relies on the corrected editing distance (μ_{ED}), for a method to obtain a good cluster homogeneity measure (μ_H), it also has to show a good savings value (good μ_{ED}).

2.5 Experimental setup

In this section, we describe the principles (relations) used for clustering (§2.5.1), the different term representations adopted for the methods evaluated (§2.5.2) and the clustering parameters for each method (§2.5.3).

2.5.1 The relations used for clustering

Given the OTC task, our experiment consisted in searching for the principle and the method that can best perform this task. Three principles were tested:

CLS: Clustering by coarse lexical similarity: grouping terms simply by identical head word. We call this "baseline" clustering as it is technically the most straightforward to implement and is also a more basic relation than the ones used by TermWatch (see §2.3.2). However, it should be noted that this head relation is not so trivial for the GENIA corpus. Indeed, Weeds et al. (2005) showed that

grouping terms by identical head words enables to form rather homogeneous clusters with regard to the GENIA taxonomy. In their experiment, out of 4,797 clusters, 4104 (85%) contained terms with the same GENIA category while 558 (12%) clusters contained terms with 2 or 3 semantic categories. A further 135 (3%) clusters contained terms with more than p semantic categories.

LSS: Clustering by fine-grained Lexico-Syntactic Similarity as implemented in the TermWatch system using the CPCL clustering algorithm described in section §2.3.3. Terms are represented as a graph of variations.

LC: Clustering by Lexical Cohesion. This principle required a spatial representation based on a vector representation of terms in the space of words they contain. It was suggested by the characteristics of the baseline and graph (LSS) representations. The LC representation offers a numerical encoding of term similarity that allows us to subject statistical clustering approaches (hierarchical and partitioning algorithms) to the OTC task. We describe this representation in more details below.

2.5.2 Vector representation for statistical clustering methods

In order for statistical clustering methods to find sufficient *co-occurrence* information in an OTC task, it was necessary to represent *term-term* similarity. We redefined *co-occurrence* here as *intra-term word co-occurrence* and built a *term* × *word* matrix where the rows were the terms and the columns the unique constituent words.

To ensure that the statistical methods will be clustering on a principle as close as possible to the LSS relations used by TermWatch and to the head relation used by the baseline, we further adapted this matrix as follows: words were assigned a weight according to their grammatical role in the term and their position with regard to the head word. Since a head word is the noun focus (the subject), it receives a weight of 1. Modifier words are assigned a weight which is the inverse of their position with regard to the head word. For instance, given the term "*coronary heart disease*", *disease* (the head word) will receive a weight of 1, heart will be weighted 1/2 and coronary 1/3.

More formally, let $W = (w_1, ..., w_N)$ be the ordered list of words occurring in the terms. A term $t = (t_1, ..., t_q)$ can be simply viewed (modulo permutations) as a list of words where the t_i are words, t_q is the head and t_1,...,t_{q-1} is a possible empty list of

modifiers. Each term t is then associated with the vector V_t such that:

$$V_t[i] = \begin{cases} \frac{1}{1+q-j} \text{ whenever } w_i = t_j \\ 0 \text{ elsewhere} \end{cases}$$

Let M be the matrix whose rows are the V_t vectors. We derive two other matrices from M:

1. a similarity matrix $S = M.M^t$ whose cells give the similarity between two terms as the scalar product of their vectors (for hierarchical algorithms).
2. a core matrix C by removing all rows of M corresponding to terms with less than three words and all columns corresponding to words that appeared in less than 5% of the terms. Indeed, experimental runs showed that the k-means algorithms could not produce meaningful clusters when considering the matrix of all terms.

This weighting scheme translates the linguistic intuition that the further a modifier word is from the head, the weaker the semantic link with the concept represented by the head. This idea shares some fundamental properties with the relations used by TermWatch for clustering. Note also that this weighting scheme is a more fine-grained principle than the one used by the baseline. Representing terms in this way leads to the identification of lexically-cohesive terms (i.e., terms that often share the same words). This idea was explored by Dobrynin et al. (2004) although in a different way. Their *contextual document clustering* method focused on the identification of words that formed *clusters of narrow scope*, i.e. lexically cohesive words which appeared with only a few other words. Lexical cohesion is not a new notion in itself. It has already been explored in NLP applications for extracting collocations (fixed expressions) from texts (Smadja, 1993; Church and Hanks, 1990).

2.5.3 Clustering parameters

MWTs were clustered following the three types of relations described in §2.5.1. The following methods were tested: baseline; CPCL on graph of variations; partitioning (k-means, Clara based on medoids), hierarchical (CPCL on similarity matrix S).

- **Baseline on CLS**: No particular parameter is necessary. All terms sharing the same head word are put in the same cluster.

- **CPCL on LSS**: Parameter setting consists in assigning a role to each relation ($COMP$ or $CLAS$). Among all the variations extracted by TermWatch, we selected a subset that optimized the number of terms over the maximal size of a class. Hence this selection was done without prior knowledge of the GENIA taxonomy. The variations selected for the $COMP$ phase are those where terms share the same head word or WordNet semantic variants. In the current experiment, by order of ascending cardinality, $COMP$ relations were:
  - spelling variants,
  - substitutions of modifiers filtered out using WordNet (sub_wn_modifier),
  - insertion of one modifier word (strong_ins),
  - addition of one modifier word to the left (strong_exp_l)
  - substitutions of the first modifier in terms of length ≥ 3 (strong_sub_modifier_3).

  The $CLAS$ variations were:
  - WordNet head substitutions (sub_head_wn),
  - insertions of more than one modifier (weak_ins),
  - addition of more than one modifier word to the left (weak_exp_l)
  - substitution of modifiers in terms of length ≥ 3 (weak_sub_modifier_3).

  No threshold was set so as not to exclude terms and relations. Since the objective of this experiment is to form clusters as close as possible to the GENIA classes, the algorithm was stopped at iteration 1. Thus, only a few part of relations induced by the variations were really used in the clustering. More precisely, only relations induced by rare variations which are assigned a higher weight or relations between near-isolated terms were considered. Hence, the exact technique used in agglomerative clustering (single, average or complete link) did not come into play here. We also tested the performance of the 1st step grouping, i.e., the level forming connected components ($COMP$) with a subset of the relations. This level is akin to baseline clustering although the relations are more fine-grained.

- **Hierarchical on LC:** Clustering is based on the similarity matrix $S[S \geq th]$ derived from S by setting to 0 all values under a threshold th. We used the following values for th:
  - 0.5 : the rationale is that at this weight, terms either share the same head or have common modifiers close to the head,
  - 0.8 : this weight imposes the same head on related terms,

Because the dissimilarity matrix was too large, we had to use our own PERL programs to handle such sparse matrices. Based on a graph representation of the data, only non zero values were stored as edge values enabling each iteration to be done in a single search. We were thus able to run the usual variants of single, average and complete link hierarchical clustering on this system but they did not produce any relevant clustering (all the cluster evaluation measures were negative). Since the similarity matrix S had all the requirements to be an input to the CPCL algorithm, we subjected it to the CPCL algorithm. After some tests, we finally selected the *vertex-weight* (§2.3.3) as the agglomerative criterion since it significantly reduced the chain effect. We did four iterations for each threshold value. This yielded significant results. Thus the results shown for hierarchical clustering were obtained using the CPCL algorithm on the *term* × *word* matrix.

- **Partitioning on LC:** This method is based on the computation of k-means centers and medoids on the core matrix C. We used the standard functions of k-means and CLARA (Clustering LARge Applications) fully described in Kaufman and Rousseeuw (1990). CLARA considers samples of datasets of fixed size on which it finds k medoids using PAM algorithm (Partitioning Around Medoids) and selects the results that induce the best partition on the whole dataset. PAM is supposed to be a more robust version of k-means because it minimizes a sum of dissimilarities instead of a set of distances. However, for large datasets, PAM cannot be directly applied since it requires a lot of computation time. CLARA and PAM are available on the standard R cluster package[6]. To initialize CLARA, we used the same procedure as CLARANS (Ng and Han, 2002) to draw random samples using PERL programs and a graph data structure. We ran these two variants (k-means and CLARA) for the following values of k: 36, 100, 300, 600 and 900. Then, given these centers and medoids, we again used our PERL programs for storing large sparse matrix, to assign each term to its nearest center or medoid and to obtain a partition on the whole set of terms.

The results of clustering with these algorithms and their variants were then evaluated against the target partition (the GENIA taxonomy) using the measures described in §2.4.3. Combining R and PERL 5 has been quite efficient. R offers very robust implementations of spatial clustering algorithms while PERL allows one to easily define optimal data structures. Thus all the data processing including the initialization phase and sample extraction was done with PERL, leaving to R the massive numerical computations based on C and FORTRAN subroutines. All the tests were performed on

[6]Version 1.10.2, 2005-08-31, by Martin Maechler, based on S original by Peter Rousseeuw (rousse@uia.ua.ac.be), Anja.Struyf@uia.ua.ac.be and Mia.Hubert@uia.ua.ac.be.

a PENTIUM IV PC server running LINUX DEBIAN stable with 1Go of RAM, SCSI disk and no X11 server for memory saving.

2.6 Results

2.6.1 Possible impact of the variations on TermWatch's performance

Before comparing the clustering results obtained by the different methods, we investigated the possible impact of the variations used by TermWatch on its performance. The idea was to determine if our variation relations alone could reproduce these categories, i.e., if they grouped together terms from one only GENIA class. In this case, then there would be no need to perform clustering since the variation relations alone can discover the ideal partition. However, our study showed that this was not the case.

The following chart figure (2.3) shows for each of our variation relation, the number of links acquired, the proportion of intra-category links and the proportion inter-category links (from different classes). We can see clearly from this figure that some relations are rare, i.e., they capture too few links although they link terms from the same class (*sub_modifier_wn, strong ins, weak ins*). These relations are in the minority especially by the proportion of terms linked. Other relations like *weak_exp2, weak_sub_head3, weak_exp_r* are more abundant but they lead to heterogeneous clusters, they link terms from different GENIA classes. Surprisingly, *weak_exp_l* and *strong_sub_mod3* produced relatively good quality clusters while relating a considerable number of terms.

2.6.2 Evaluation of clustering results

Using the relations chosen in §2.5.3, CPCL on LSS generated $1,897$ non trivial components (at the COMP phase) involving only $6,555$ terms. Adding CLAS relations in the second phase led to $3,738$ clusters involving $19,887$ terms.

Hierarchical clustering based on similarity matrix introduced in §2.5.2 generated $1,090$ clusters involving 25.129 terms for a threshold $th = 0.5$ and $1,217$ clusters involving $19,867$ terms for $th = 0.8$.

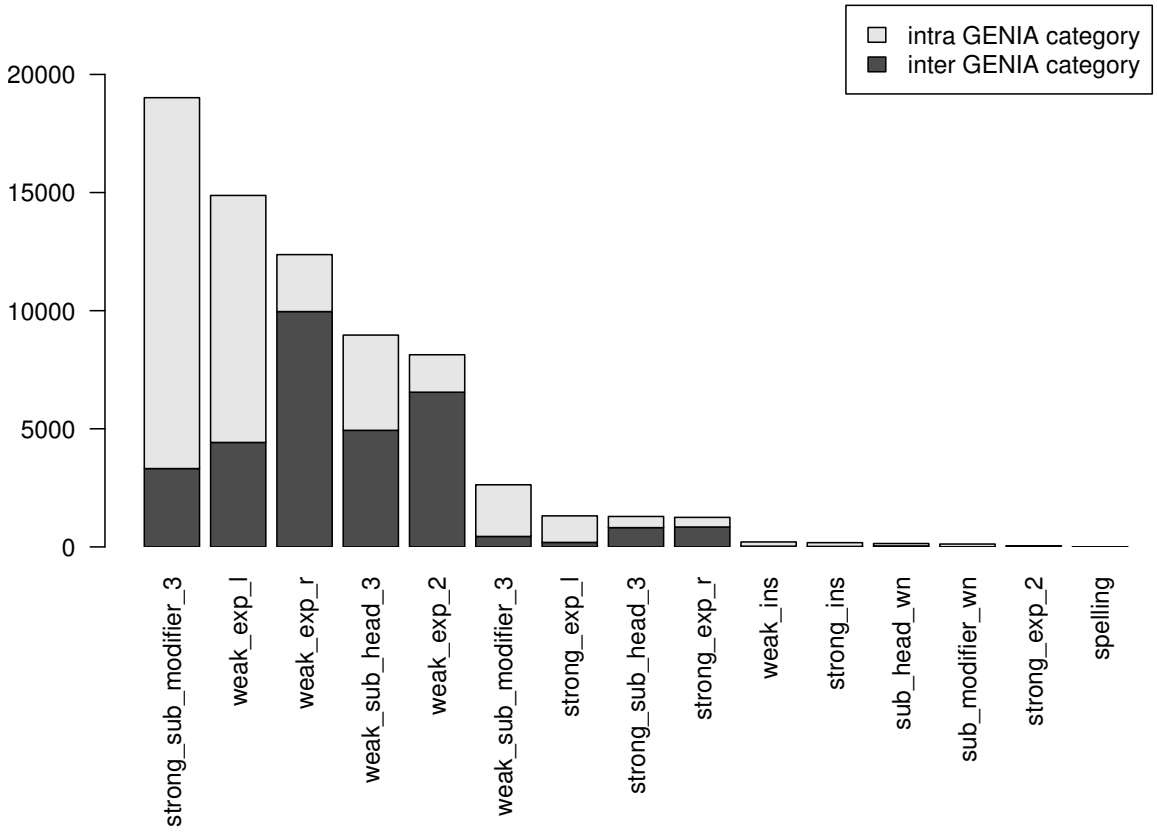

Figure 2.3: Distribution of related pairs of terms by variations.

The plots in figures 2.4 and 2.5 show the results of the evaluation measures μ_{ED} and μ_H introduced in §2.4.3. Since the majority of the clustering methods are sensitive to term length, we plotted the score obtained by each of the measure (y-axis) by term length (x-axis). Note that at each length, only terms of that length and above are considered. For instance, at length 1, all terms are considered. At length 2, only terms having at least two words are considered. Thus, the further we move down the x-axis, the fewer the input terms for clustering.

Figure 2.4 shows the % of savings obtained by the nine algorithms tested using the corrected ED measure. We see that the hierarchical method with a threshold = 0.8 and CPCL obtain a better score than the baseline clustering when considering all the terms (length ≥ 1). When fewer and longer terms are considered (length ≥ 3), partitioning methods outperform CPCL and hierarchical algorithms but still remain below the baseline. This is because, at length ≥ 3, CPCL has fewer terms, thus fewer relations with which to perform the clustering. Statistical methods on the other hand, with longer terms have a better context, thus more relations in the matrix. From terms of length ≥ 4 words, partitioning methods outperform the baseline.

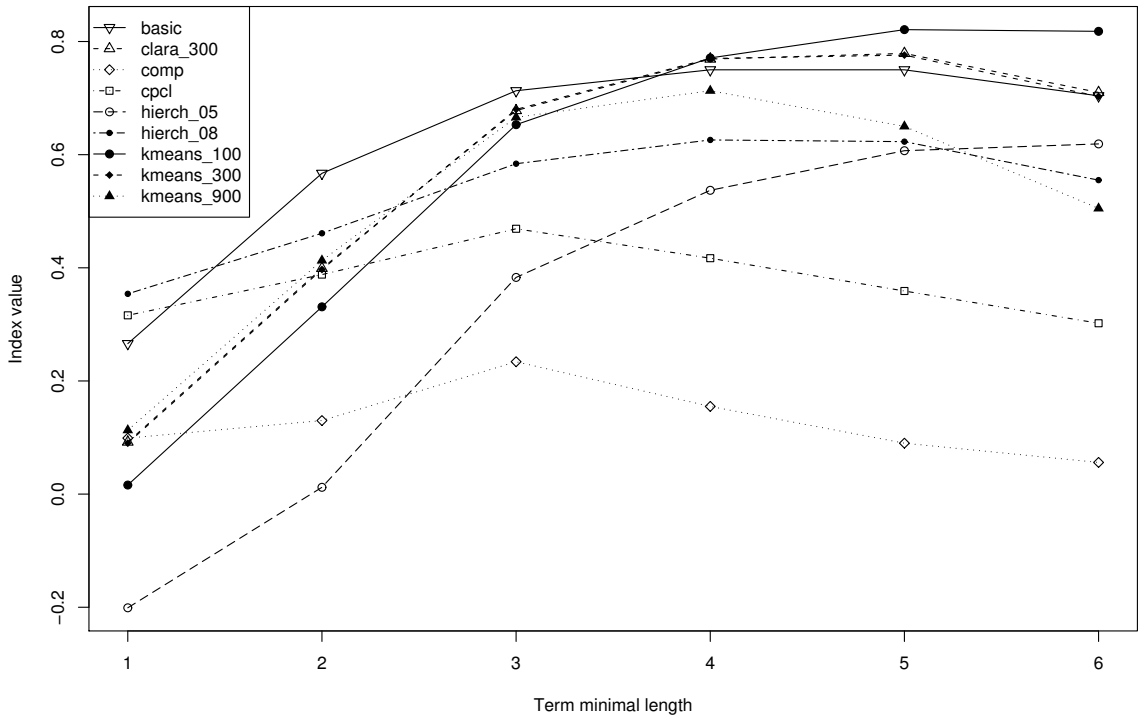

Figure 2.4: Editing distance between clustering results μ_{ED} and Genia categories.

However, the ED measure masks important features of the clustering outputs since it is a compromise between the number of necessary moves and merges needed to reach the target partition. More important is the quality of the clusters (cluster homogeneity) vis-à-vis the target partition (GENIA classes). This is measured by the μ_H which calculates the ratio between the value of ED and the number of movings. The μ_H performance of the algorithms is shown in the plot of figure 2.5.

It appears clearly that on cluster quality, CPCL is the only algorithm that significantly outperforms the baseline irrespective of term length. Hierarchical algorithm with $th = 0.8$ and the COMP phase of CPCL follow closely but only on all terms (length ≥ 1). Their performance drops when terms of length >= 4 are considered. Partitioning algorithms show poor cluster homogeneity. K-means with $k = 100$ performs worse than the other variants (Clara, k-means with $k = 300$, $k = 900$). Hierarchical with $th = 0,5$ obtain the poorest score.

To gain a better insight on the cluster homogeneity property, we generated for every algorithm a chart showing the proportion of terms which share the same GENIA class with the majority of terms in the same cluster (and thus that do not require any move) The charts in figures 2.6-2.9 show the proportion of intra and inter GENIA category

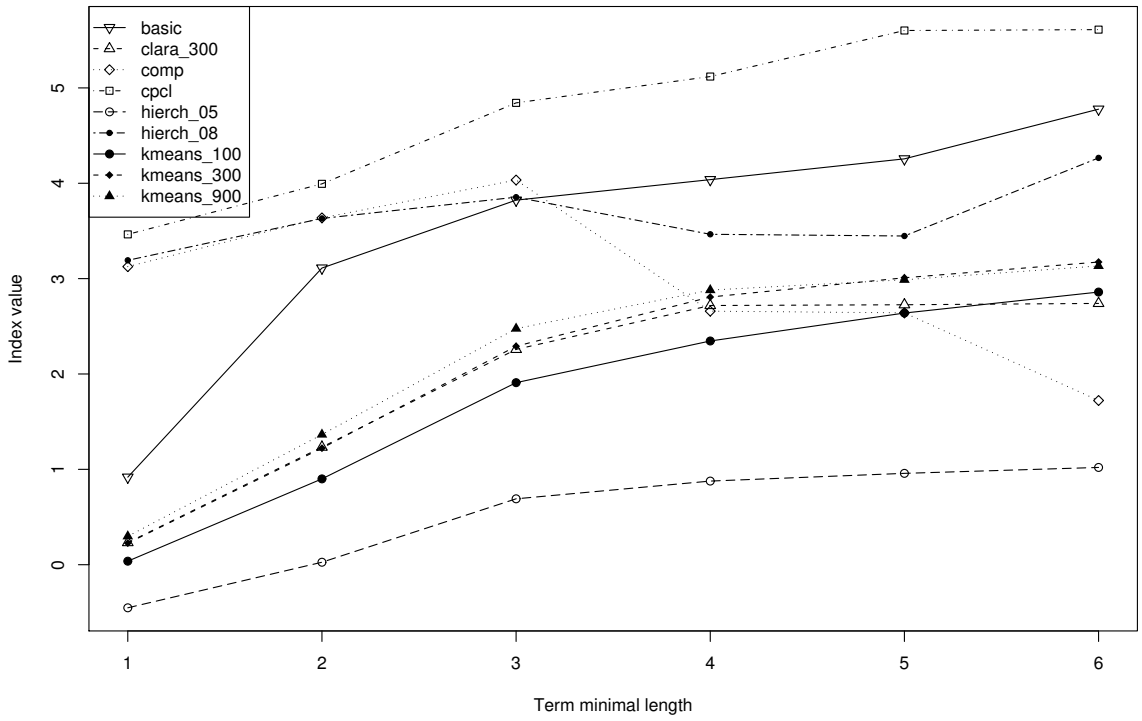

Figure 2.5: Cluster homogeneity measure μ_{ED} on the Genia categories.

terms for single link clustering algorithms. The black bars represent mis-classifications. White bars represent terms from the same GENIA category.

It appears that the COMP variant of CPCL produced the most homogeneous clusters which is not altogether surprising because the relations used in COMP phase are the most semantically tight. COMP and CPCL significantly outperform the baseline. This good performance is a bit unexpected for CPCL because the CLAS relations induce a change of head word which could lead to a semantic gap (change of semantic class).

Closely following is the hierarchical algorithm at $th = 0.8$ denoted “hierch_08" in the figures.

The baseline comes fourth which shows that grouping terms simply by identical head words as done by baseline is good but not good enough to form semantically homogeneous clusters.

Partitioning methods produced less homogeneous clusters with low error rates roughly on categories with a low proportion of one word terms.

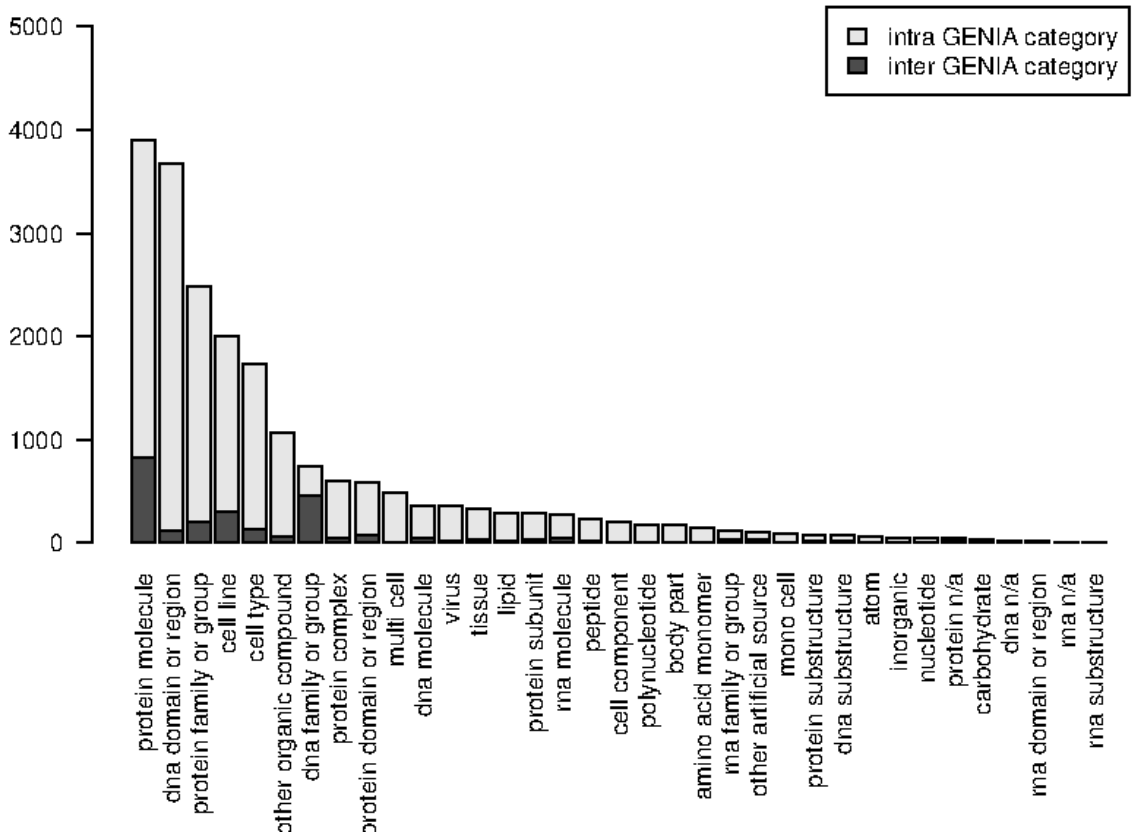

Figure 2.6: CPCL clustering results.

2.7 Concluding remarks

We have developed an efficient text mining system based on meaningful linguistic relations which works well on MWTs and thus on very large and sparse matrices. This method is suitable for highlighting rare phenomena which may correspond to weak signals.

The specific evaluation framework set up here led us to redefine a matrix representation in order to enable comparison with existing statistical methods. We defined a new term weighting scheme in the matrix representation enabling statistical methods to build significant clusters. We also corrected an existing cluster evaluation measure and defined a complementary one focused on cluster homogeneity.

The choice of the evaluation metric made it possible to compare algorithms outputting very high number of clusters, with considerable differences in this number (between 100 for K-means and 3,738 for CPCL). This was done without any assumption of

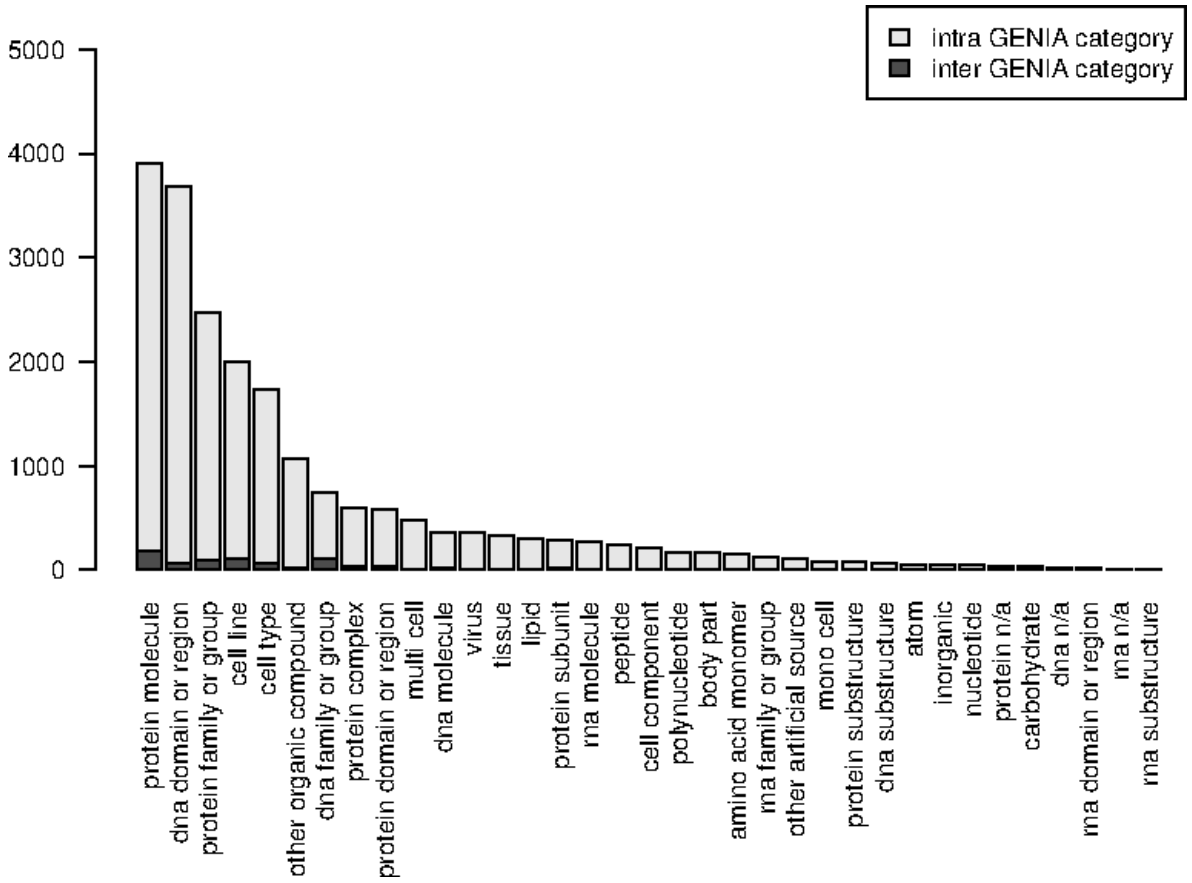

Figure 2.7: COMP clustering results.

equal cluster size. We believe these differences did not handicap any algorithm unduly since all produced clusters whose numbers were very far from the target partition (36 classes), especially our own method. As we cannot define a priori the number of optimal clusters, CPCL's performance was hampered for the μ_{ED} measure. Statistical methods (both hierarchical and partitioning) were more sensitive to term length.

The results however show that CPCL performs well in terms of cluster quality (homogeneity). Since this approach is computationally tractable in linear time, it also appears to be the best candidate for tasks requiring interaction with users in real time, like interactive query refinement. This aspect will be explored in a separate study.

Overall, this experiment has shown that even without adequate context (document co-occurrence), clustering algorithms can be adapted to partially reflect a human semantic categorization of scientific terms.

Another interesting finding of this study is that when considering an OTC or a similar task, it may be interesting to first consider clustering by a basic relation before

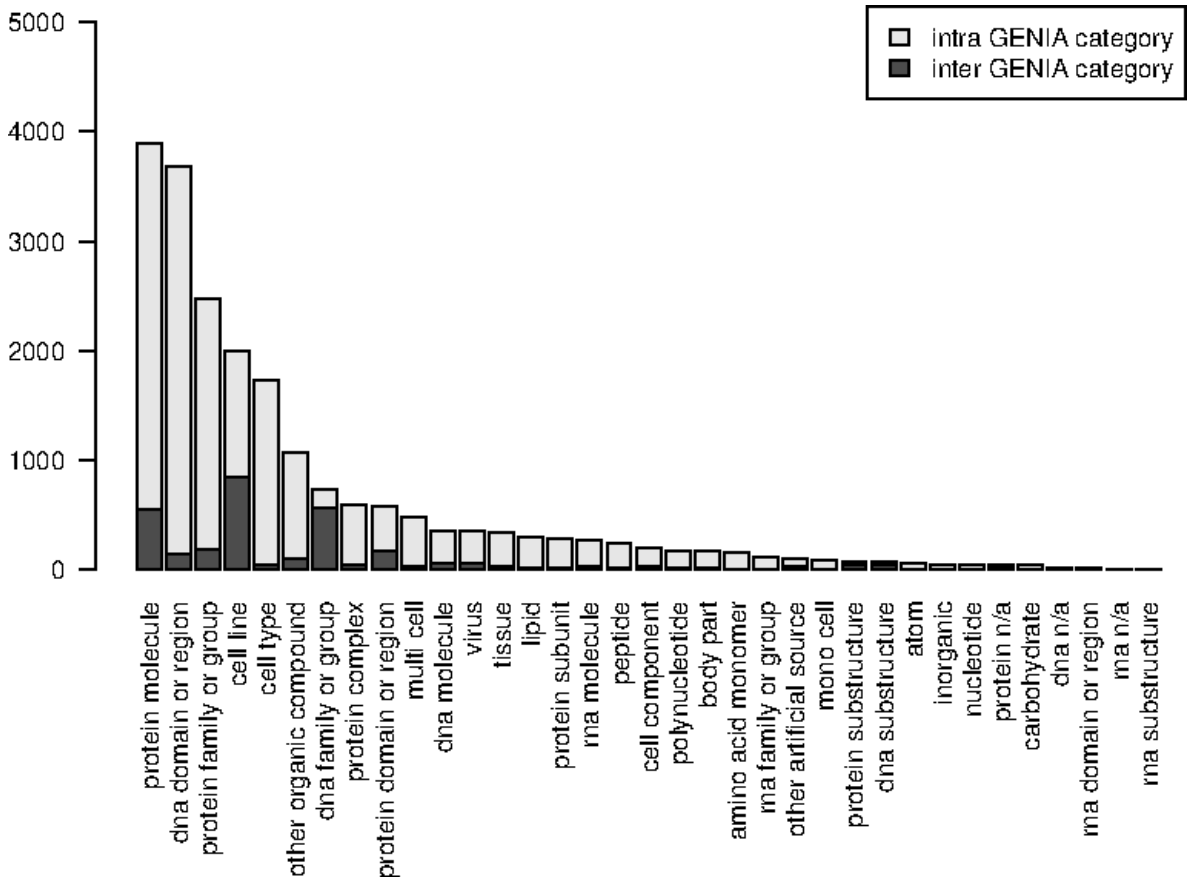

Figure 2.8: Hierarchical clustering clustering results.

resorting to more complex and fine-grained term representation. The performance of the baseline clustering in our experiment is far from poor. It could be satisfactory for some tasks, for instance as a first stage for learning new taxonomy or knowledge structures from texts. These can be further refined using more sophisticated approaches: fine-grained linguistic relations, machine learning techniques with manually tagged learning sets.

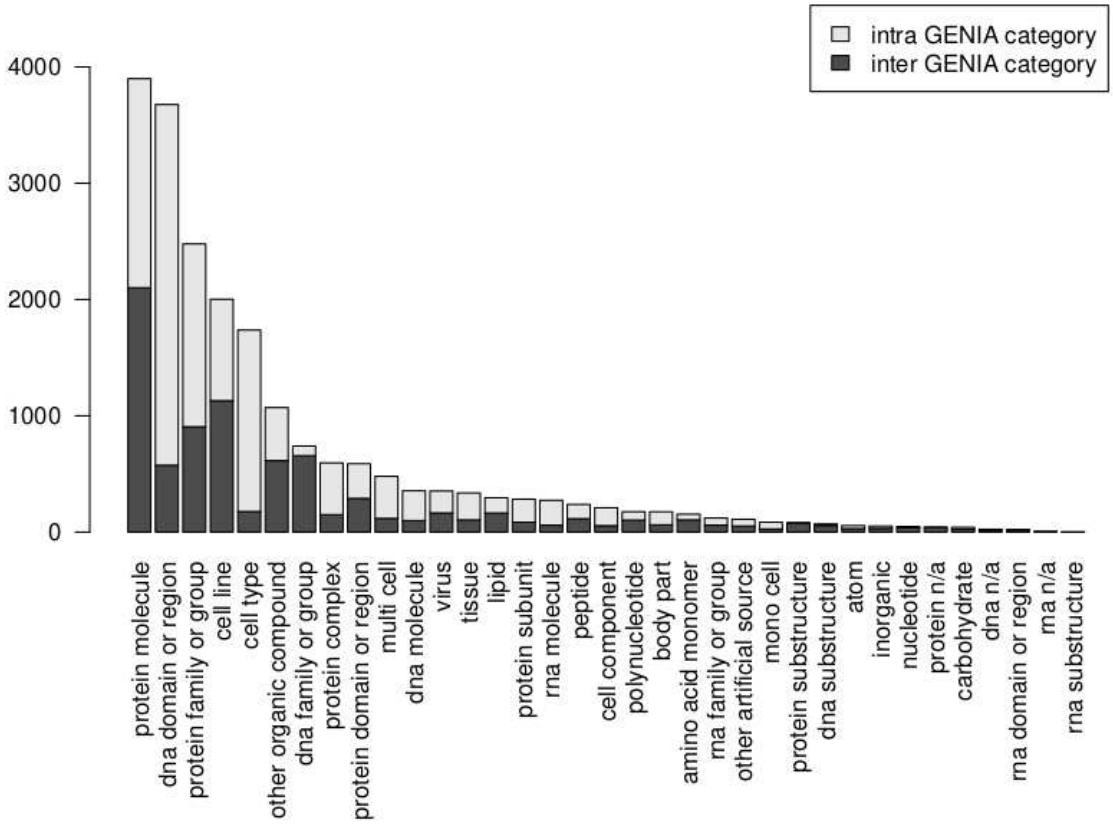

Figure 2.9: Baseline clustering results.

Chapter 3

Mapping knowledge by automatic extraction of terminology graphs

3.1 Introduction

A timely awareness of recent trends in scientific domains is necessary to support several information intensive activities such as innovation, science and technology watch, business intelligence to name only a few. Such studies are usually conducted by analyzing the electronic literature available on line based on different approaches such as citation analysis, text and document clustering, pattern mining, novelty detection. Bibliometrics aims to elaborate indicators of the evolution of scientific activities using statistical and mathematical models. The two major bibliometric methods are co-citation and co-word analysis. Co-citation analysis has proved useful in highlighting major actors in a field (the "who's who" of a field). Although some attempts have been made to work directly at the text level in bibliometrics, natural language processing (NLP) resources and capabilities have barely been tapped by this community. The most common NLP processing is limited to stemming Porter (2006) prior to clustering Zitt and Bassecoulard (1994); Glenisson et al. (2005). Text units have mainly been considered either as a bag-of-words or as a sequence of n-grams in the vast majority of topic mapping systems.

We take a different approach to text clustering and consider that a multi-disciplinary effort integrating surface linguistic techniques is necessary to elaborate indicators of topics trends at the level of texts. For this, we require a more fine-grained analysis, involving prior linguistic processing of the scientific literatures before applying statis-

tical and mathematical models. The interesting features of our approach lie in the combination of state-of-the-art techniques from three disciplines: Natural Language Processing (NLP), Data Mining and Graph Theory. NLP enables us to extract meaningful textual units and identify relevant information between them, here multi-word terminological units. These text chunks correspond to domain concepts and the linguistic relations are lexical, syntactic and semantic variations. These variations are used in later stages of processing (clustering) to form topics through relations of synonymy and hyponymy/hypernymy and semantic relatedness. Prior grouping of term variants ensures that semantically close terms which reflect different aspects of the same topic are certain to end up in the same cluster at the end of the process. The linguistic theory behind the grouping of terms either by shared modifiers or by shared head is known as distributional analysis and was introduced by Harris Harris (1968). It was later taken up by various studies in automatic thesaurus construction Grefenstette (1997); Watcholder et al. (2001). Ibekwe-Sanjuan Ibekwe-SanJuan (1998b) extended the types of identified relations and defined additional constraints like the position of added words and their number to avoid generating spurious variants. This approach has been implemented in the TermWatch system Ibekwe-SanJuan and SanJuan (2003, 2004); SanJuan and Ibekwe-Sanjuan (2006). There co-occurrence (numerical) is optionally added during clustering as a means to capture the supplementary dimension of interactions between domain concepts. The end results are clusters of high semantic homogeneity which also capture the most salient association links.

TermWatch implements a hierarchical clustering algorithm to suit the characteristics of multi-word terms. This algorithm clusters the multi-word terms grouped into close semantic classes called components using optionally co-occurrence information. The clusters are represented as an undirected graph. The system has been applied successfully to text corpora from different domains and on several knowledge intensive tasks such as knowledge domain mapping in bio-technology Ibekwe-SanJuan and Dubois (2002), ontology population in the biomedical domain Sanjuan et al. (2005), opinion categorization of literature reviews Chen et al. (2006).

Here we present an enhancement to the system by integrating a graph decomposition algorithm studied in Biha et al. (2007) which enables the system to decompose complex graphs into more legible subgraphs representing coherent networks of research topics. This allows to split complex terminological networks of topics extracted by TermWatch based on their graph theoretic properties in order to identify sub-structures that represent highly connected sets of topics called central atom and distinct sets of topics called peripheral atoms.

In Ibekwe-Sanjuan et al. (2008), we have applied previous versions of TermWatch

II with an earlier implementation of the graph decomposition algorithm to scientific publications related to the Sloan Digital Sky Survey[1]. Here, we apply the to mapping knowledge in terrorism research between 1990-2006. The datasets are publication records of peer-reviewed journal articles downloaded from the Web of Science (WoS). The input to our system are the titles and abstract, publication year and author fields of the records. We favored using the WoS database as it indexes high quality journals with high impact factor in their respective fields.

Since the sept 9/11 attack, a lot of attention has been focused on rapidly detecting indicators of potential terrorist threats. This corpus was built following a search on the WoS using the word "terrorism". 3,366 bibliographic records were collected. Note that this corpus is not on individuals or groups involved in terrorist acts but rather on what researchers have been writing about terrorism: its effects on the victims and the general public, its forms, its means and ways to prepare for it. Previous studies have sought to map the terrorism domain either from this same perspective Chen (2006) or from that of groups actively involved in plotting and carrying out terrorist acts Chen et al. (2008). Of particular relevance to our study is the one done by Chen (2006). This author used the same database and the same query but on an earlier and shorter period (1990-2003). His results, validated by domain experts will serve as a "baseline" against which we compare our system's performance. Our analysis of the terrorism research dynamics is thus a follow-up of his study but using a different methodological approach.

The rest of the chapter is structured as follows: Section 3.2 is a general description of TermWatch. Section 3.3 details the terminological graph extraction process. We then show in section 3.4 how an association graph can highlight a family of formal concepts and their relations based on the unique atom decomposition. Section 3.5 analyzes results obtained from the two case studies and section 3.6 draws some conclusions from this experiment.

3.2 Overview of TermWatch

TermWatch is designed to map research topics from unstructured texts and track their evolution in time. The system combines linguistic relations with co-occurrence information in order to capture all possible dimensions of the relations between domain concepts. It is currently run on-line on a LINUX server[2]. Standalone terminology graph construction and decomposition modules are available under the GNU public license

[1] http://www.sdss.org/

[2] http://system.termwatch.es

(GPL). The processing of texts relies on surface linguistic relations between multi-word terms (MWTs) to build semantically tight clusters of topics. The processes leading from raw texts to the mapping of domain topics can be broken down into five major stages: multi-word term extraction, term variants identification, term clustering, graph decomposition and visualization. Figure 3.1 shows the overall process. As some components of the system have been described in previous publications Ibekwe-SanJuan and SanJuan (2003, 2004); SanJuan and Ibekwe-Sanjuan (2006), we will focus particularly on the graph decomposition algorithm of terminological graphs which aims to reveal a family of formal concepts and their relationships. A step-by-step procedure going from input texts to topic mapping consists in the following:

1. Build a scientific corpus reflecting a research question. The input corpus is composed of raw texts.

2. Terminological noun phrases (NPs) of maximal length are extracted using TreeTagger Schmid (1994) or any POS tagger. A selection of NPs is done based on their syntactic structure and on our enhanced term weighting function in order to retain only domain terms.

3. Terms that are semantic variants of one another are detected and clustered in a hierarchical process. This results in a three level structuring of domain terms. The first level are the terms. The second level are components that group together terms semantically close terms or synonyms. Roughly, TermWatch's components generalize the notion of WordNet synsets Miller (1994) to multi-word terms. A clustering algorithm Ibekwe-SanJuan (1998a) is applied to this second level of term grouping based on a weighted graph of term variants. Components and clusters are labeled by their most active term and can be used as document features.

4. In the fourth stage, documents are indexed by cluster or component labels and the corresponding association graph is generated. The strength of the association is weighted based on different similarity measures and only those links that are above some threshold for all measures are considered.

5. Association graphs are decomposed into atoms Biha et al. (2007). An atom is a subgraph without clique separators. Each clique corresponds to a formal concept. Major atoms are detected and visualized using force directed placement algorithms. The periphery of big atoms is highlighted since it can reveal new concepts arising in a domain represented by a central more bigger atom.

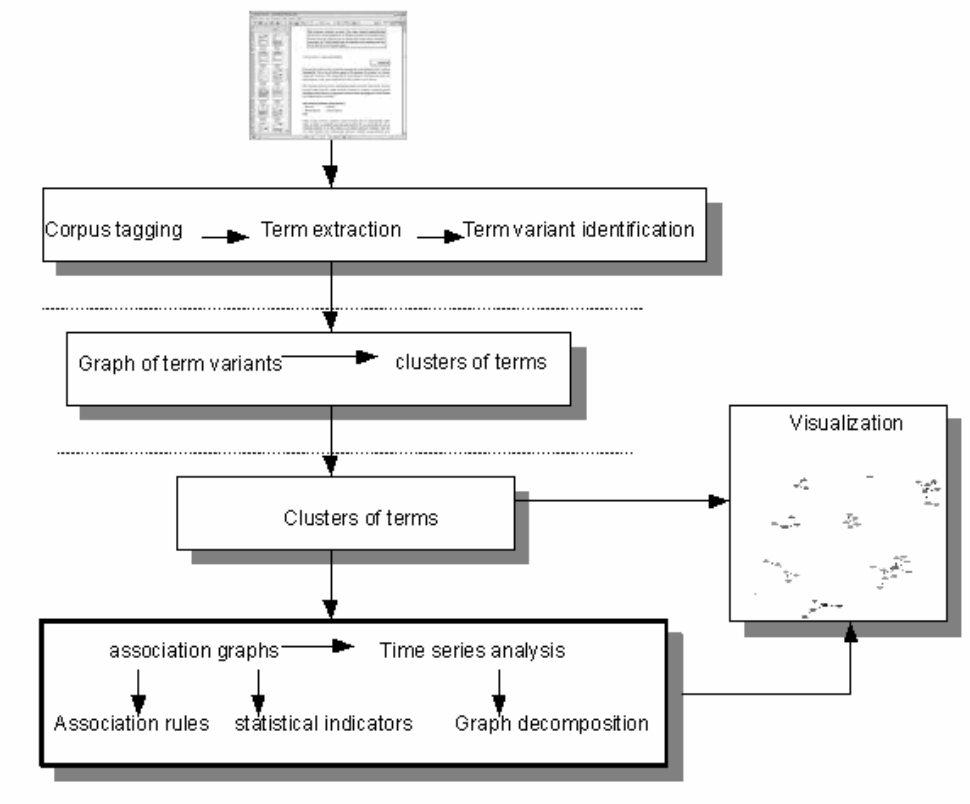

Figure 3.1: Overview of the mapping knowledge domains process in TermWatch II

3.3 Terminological graph extraction

3.3.1 Term Extraction

After the corpus has been tagged using TreeTagger Schmid (1994), contextual rules are used to extract multi-word terms based on morphological and syntactic properties of terms. One such rule is shown in fig. 3.2. This rule favors the extraction of terminological noun phrases in a preposition structure where the preposition is “of". This preposition has been found to play an active role in the multi-word term formation process. More details of the rules can be found in Ibekwe-SanJuan (1998b). The extracted terms can be simplex noun phrases (NPs) like “stress disorder" or complex ones like “posttraumatic stress disorder" which embeds simpler NPs. Also, terms are extracted in their two possible syntactic structures: NPs with prepositional attachment (execution of innocent victims) and compounds (innocent victims execution). This transformation operation, also known as permutation is useful for grouping together syntactic variants

of the same concept that would otherwise be dispersed. No limit is imposed on the length of the extracted terms thus ensuring that new terms coined by authors of papers are extracted 'as is' and that existing domain concepts with multi-words are not altered or lost. By not resorting to the usual "bag-of-word" approach common in the IR and data mining communities, emergent domain terms can be identified in a timely manner because term extraction respects the structure of the domain terminology "in-the-making".

If $< mod >^{\star}< N >^{+}$of$< mod >^{\star}< N > + < prep1 >< verb >< mod >^{\star}< N >^{+}$
then return: $< mod >^{\star}< N >^{+}$ *of* $< mod >^{\star}< N >^{+}$ **and** $< mod >^{*}< N >^{+}$ **where:**
$< mod >$ is a determiner or an adjective
$< N >$ is any of the noun tags
$< prep1 >$ is all the prepositions excluding "of"
* is the Kleene's operator (zero or n occurrences of an item)
$^{+}$ is at least one occurrence

Figure 3.2: Example of contextual rules used to extract multi-word terms

3.3.2 Generating a graph of semantic term variants

We studied linguistic operations between terms which are domain independent and can be used to build taxonomies, thesaurus or ontologies. These operations, called terminological variations, stem from two main linguistic operations: lexical inclusion and lexical substitution. By lexical inclusion, we refer to the case where a shorter term is embedded in a longer one through three specific operations: insertions (severe poisoning $\leftrightarrow$ severe food poisoning), modifier or head word expansion ("disaster intervention" $\leftrightarrow$ "disaster intervention call"). By lexical substitution, we refer to the case where terms of identical length share a subset of lexical items save one in the same position ("political violence threat" $\leftrightarrow$ "political violence campaign"). Lexical inclusion often engenders hypernym/hyponym (generic/specific) relations between terms while the lexical substitution tend to indicate a loose kind of semantic association between terms. Lexical substitutions between binary terms give rise to a highly connected graph of term variants (cliques) which may include some amount of noise (spurious relations). They are filtered using two criteria: we retain only those substitutions that involve terms of length ≤ 2 if the words in the same grammatical position are found in the same WordNet synset. Although there are many more types of linguistic relations, we restricted our choice to those that did not require heavy use of external semantic resources and were domain-independent, thus found in any well written text revolving around the same broad topic.

We also acquired explicit synonymy links between multi-word terms using WordNet. To do this, we extended the single word-word relations in WordNet to multi-word terms by adding these restrictions: two multi-word terms are considered to be in a synonymy relation if two of their words are in the same WordNet synset, occupy the same grammatical role in the terms (both head words or modifier words) and are found in the same position. Table 3.1 shows some of the synonyms identified in this way. The italicized words were in the same WordNet synset.

Term	**Synonym identified using WordNet synsets**
september 11 *wake*	september 11 *aftermath*
united states federal *agency*	united states federal *bureau*
risk society *conception*	risk society *concept*
Trauma type	*injury* type
Life-threatening problem	Serious problem
Cyber-terrorist attack	*hacker* attack

Table 3.1: Some synonyms acquired from the terrorism corpus using WordNet synsets.

Table 3.1 shows that the quality of the synonyms acquired through WordNet is indeed good. Explicit synonymy links ensure that concepts appearing under different names are not dispersed in different clusters at the end of the process. Table 3.2 gives examples of the different relations identified and the number of terms involved for the terrorism corpus.

Variation type	**example of Term**	**example of Variant**	**#Terms**	**#Links**
Spelling	trauma *center*	trauma *centre*	93	138
Left exp.	food contamination	*pet* food contamination	1,799	2,709
Insertion	poisoning case	poisoning *medical intervention* case	41	60
Right exp.	disaster intervention	disaster intervention *call*	2,884	4,326
Modifier sub.	*acute* stress disorder	*posttraumatic* stress disorder	14,062	95,651
Head sub.	political violence *threat*	political violence *campaign*	13,810	125,385
Wordnet Mod. sub.	*Trauma* severity	*injury* severity	185	99
Wordnet Head sub.	terrorist *financing*	terrorist *funding*	396	217

Table 3.2: Terminological variations identified between terms in the terrorism corpus.

Any relation between a set of documents and a set of features naturally induces a network of associations. Two features are associated if they index a substantial set of common documents. The association can therefore be weighted by a measure on the set of shared documents. The network of associations gives rise to a *feature* $\times$ *feature* symmetric matrix that can be analyzed using standard data mining approaches like clustering, factor analysis or latent semantic analysis. The output of these methods heavily depends on the choice of the association index. However, before applying

any data mining process, the structure of the association network should be studied independently from the measure of associations.

The study of this structure becomes indispensable when features result from a complex text analysis process like multi-word terms (MWTs) extracted from abstracts in an automated procedure. Since these terms result from an unsupervised process, some amount of noise can be expected. The idea is then to use standard association measures to remove the most improbable associations. So, instead of working on a numeric matrix, we consider the binary matrix that indicates if an association between two multi-word terms is possible or not, without prejudice on its strength since it could result from some bias in the term selection procedure. Moreover, low frequency terms are essential when seeking for rare information like emerging new concepts and/or new relationships between concepts. This symmetric binary matrix gives rise to a non directed graph between multi-word terms. In the case of a corpus of documents constituted randomly, the structure of this graph corresponds to the usual small world frequently observed on co-word graphs Ferrer i Cancho and Solé (2001). In some cases, the extracted terminological network of possible associations shows an unexpected structure. TermWatch II aims to extract terminological graphs and to reveal this structure if it exists, based on advanced graph algorithm theory.

3.3.3 Term Clustering

The linguistic significance of each relation can be translated in terms of two possible roles: COMP and CLAS. Ideally, COMP relations are variations that induce near-semantic equivalence or synonymy links such as spelling variants, permutations, WordNet synonyms, one-word modifier expansions and insertions. COMP relations are used to form a prior category of tight semantic clusters which serve as a first level of agglomeration. There is an edge between two nodes if one is a COMP variant of the other. By forming connected components, we group terms for which there is a sequence of variations in COMP. Since variations in COMP link only close semantically related terms, resulting connected components portray terms from the same concept family. Components are labeled by its most central term and can be used as document descriptors. CLAS relations are those that involve a topical shift between two terms, i.e., where the head word is different like head expansion and head substitution. For instance, the shift of focus from “criminal assault” to the victim in “criminal assault victim”. This category of relations is used to aggregate the components formed by COMP relations in an agglomerative hierarchical process.

The strength of these links between components can be measured by the number

of variations across them. In other to favor rare relations and eliminate noise, each variation is weighted by the inverse of its frequency in the corpus. Then the strength of the link between two components I, J is computed as follows:

$$d(I,J) \quad = \quad \sum_{\theta \in CLAS} \frac{N_\theta(I,J)}{|\theta|} \tag{3.1}$$

where $N_\theta(I,J)$ is the number of variations of type θ in a subset of relations not in COMP called CLAS ($CLAS \cap COMP = \emptyset$) that relate terms in I to terms in J. $|\theta|$ is the total number of variations in θ.

CLAS clusters can be then formed using any graph clustering algorithm based on this valued graph of components. TermWatch implemented CPCL (Clustering by Preferential Clustered Link) algorithm, first described in Ibekwe-SanJuan (1998a). The principle of CPCL algorithm is to select at each iteration edges that are local maximums and merge iteratively together all nodes related by such edges. The advantage of this principle is that two nodes are merged not only based on the strength of their relation but also by considering all the relations in their neighborhood. The system then merges the components with the strongest relation at iteration t. We have shown in SanJuan and Ibekwe-Sanjuan (2006) that CPCL has a unique possible output and avoids part of the chain effect common to similar hierarchical clustering methods. CPCL is also different from usual hierarchical clustering (single, average, complete link) since more than one group of components can be clustered at different similarity values. We refer the reader to SanJuan and Ibekwe-Sanjuan (2006) for a more formal description as well as for a comparison with a larger family of clustering algorithms (variants of single-link, average link and variants of k-means). Table 3.3 shows as example, the content of the biggest component in the biggest cluster. This cluster has 78 terms and has been automatically labeled "terrorist attack" which is the label of its major component. The other terms in the cluster result from co-occurrence links. We also show in the lower part of this table, surrounding nodes around this cluster which form a clique.

3.4 Association Graph analysis

Clustering a large corpus of terms can lead to several hundreds even if coherent clusters which are difficult to visualize (cluttered image). We also need to study the way in which these clusters are associated to documents.

Terms in component “terrorist attack"
terrorist attack, presumed terrorist attack, limited terrorist attack, national terrorist attack, international terrorist attack, explosive terrorist attack, deliberate terrorist attack, deliberate smallpox terrorist attack, smallpox attack, covert smallpox attack, chemical terrorist attack, th terrorist attack, year terrorist attack
Some components in the clique around “terrorist attack"
anthrax infection, toxic chemical, medium representation, 9/11 event, september 11 attack, current PTSD, new york time, pharmaceutical industry, american history, united kingdom, potential terrorist, militant islam, safety sense, national terrorist attack impact, distress symptom, decontamination area, immigration policy

Table 3.3: Main component of the cluster “terrorist attack” and related clusters.

3.4.1 Generating association graphs and formal concepts

In the context of association mining as defined by Agrawal et al. Agrawal et al. (1993) each document is related to the clusters that contain at least one term in the document. Clusters are then considered as items and each document defines an itemset. We shall call them document itemsets. The set of items can be extended to other fields (features) like authors. Given an integer threshold S, a frequent itemset is a set of items that are included in at least S document itemsets. There is no fixed size for frequent itemsets. Frequent itemset discovery in a data base allows to reveal hidden dependences in general. Frequent itemsets of size one are just frequent terms or authors. Frequent itemsets of size 2 induce an association graph where nodes are items and there is a link between two nodes i and j if the pair $\{i, j\}$ is a frequent itemset. Moreover, any frequent itemset defines a clique in the original association graph. Clearly, if $I = \{i_1, ..., i_n\}$ is a frequent itemset, then any pair i_k, i_p of elements in I is a frequent itemset of size two and defines an edge in the association graph but not necessarily on the graph of selected edges using a relevance measure. Thus all nodes $i_1, ..., i_n$ are related in the original association graph. However, not every clique in a graph induces a frequent itemset.

The resulting association graph being generally too dense to be visualized, it is usual to perform feature selection based on some measures like mutual information or log likelihood, to select most relevant edges. This approach has two drawbacks. First, the resulting graph structure depends on the selected measures. Second, it is not adapted to highlight larger itemsets (triplets or more). Therefore, to visualize large frequent itemsets on the association graph, we need a decomposition approach that preserves cliques induced by frequent itemsets.

The theoretical framework of association discovery is Formal Concept Analysis

(FCA) Ganter et al. (2005) based on Galois lattice theory. FCA offers a pragmatic way of formalizing the notion of concepts. It posits that to every real concept in a domain corresponds a formal concept in some database of specialized documents. In the present context, a formal concept consists of an extension made of a set D of documents, and an intension made of a set of items I such that a document d is related to all items in I if and only if d is in D. Thus a formal concept establishes an exact correspondence between a set of documents and a set of items. Frequent itemsets that are the intensions of some formal concept are called closed itemsets. We shall focus on graph decomposition methods that preserve the cliques induced by closed frequent itemsets.

3.4.2 Graph decomposition

Algorithms to enumerate all closed frequent itemsets are exponential because the number of these frequent itemsets can be exponential. Moreover they are highly redundant Zaki (2009). Thus, available packages to mine them like state of the art arules from the R project [3] require the analyst to fix a maximal size for mined itemsets. Interestingness measures are then applied to rank them. However, the list of top ranked frequent itemsets heavily depends on the choice of this measure.

Our idea is to apply the results from recent research on graph theory Berry et al. (2010); Biha et al. (2007) to extract sub-graphs that preserve special cliques that have a high probability to be closed frequent itemsets. We focus on minimal clique separators, i.e. cliques whose removal from the original graph will result in several disjoint subgraphs. This leads to extracting maximal sub-graphs without minimal clique separators. These maximal sub-graphs are called atoms Biha et al. (2007). By revealing the atomic structure of a graph we also reveal: (i) special concepts that are interfaces between sub-domains or between domain kernels and external related objects; and (ii) aggregates of intrinsically related concepts at the heart of the domain. A key point of atom decomposition is that it is unique. It is an intrinsic graph property. It follows that the number of atoms and their size distribution can be considered as good indicators of their structure complexity. Moreover the atomic structure can be computed in quadratic time on the number of nodes: $O(\#\text{vertex}.\#\text{edges})$.

In the case of mapping the structure of a domain based on a corpus of abstracts resulting from a multi-word query, it can be expected to find the concept corresponding to the query at the heart of the association graph in a central atom. This central atom

[3] http://cran.r-project.org/web/packages/arules/index.html

should contain all concepts directly related to the domain as sub-cliques. Some of them should connect the domain with external concepts and thus should be at the intersection of the central atom with peripheral ones. The atom decomposition algorithm is implemented in a C++ program Biha et al. (2007). It computes the atomic graph structure and generates two images:

- the sub-graph that constitutes the central atom if it exists.
- the network of atoms to visualize those at the periphery and the way they are connected to the central atom.

We have experimentally checked that atoms do not break 98% of closed frequent itemsets Biha et al. (2007). In the result section, we shall focus on the central atom because we found out that in the corpus analyzed here and the one in Ibekwe-Sanjuan et al. (2008), they have a surprisingly clear structure.

3.4.3 Graph visualization

The atom graphs are generated in GDL format (Sander 1995) for AiSee[4]. GDL allows to define sub-graphs objects that can be displayed folded or wrapped in a colored background. We use this functionality to fold clique sub-graphs of nodes such that the probabilities $P(i/j)$ of finding one related to a document knowing that the other is related are equal for all pair of nodes in the clique. These cliques are then represented by a generic node to simplify the display of the graph without altering its structure. We use AiSee because this software implements optimized force direct graph display algorithms Fruchterman and Reingold (1991). To analyze a complex graph structure. AiSee runs with maximal non crossing heuristics and a great number of iterations to approximate as far as possible a planar graph without crossing edges and separating non connected nodes clearly. The resulting images allow experts to quickly identify the main structural properties of the graph: maximal cycle length, connectivity, sub-cliques etc. Moreover, since nodes are labeled, domain specialists can also easily read these graphs using the browsing function of AiSee.

[4] http://www.aisee.com

3.5 Case study

We present results on mapping the dynamics of research in terrorism research between 1990-2006. An association graph between cluster labels and authors was built and subjected to the graph decomposition algorithm. The analysis of results and evaluation of the graphs is done by comparing the structure of the central atom to the network obtained by Chen (2006). His study was on the same topic, using the same query on the same database (WoS) but on an earlier period (1990-2003). Given that he has already performed an evaluation of his results by sending questionnaires to domain experts, we highlight the similarities and differences in the map he obtained and more importantly show the evolution of research on terrorism since 2003.

3.5.1 Network of atoms

The graph decomposition splits the association graph into a central and peripheral atoms. Owing to space limitations, we cannot show the images of the peripheral atoms[5]. We comment briefly on the most prominent ones. The map of atoms shows that indeed, it is a central atom on "biological terrorism" that makes the whole graph connected. Biological terrorism thus acts as a hub or a magnet for linking all the terrorism-related research. The most prominent peripheral atoms are somehow connected to this threat of bio-terrorism. The three biggest sub-graphs by number of atoms contained are "nuclear radiation" (37), "biological and chemical warfare" (25), "radiological dispersion device" (21).

3.5.2 Structure of the central atom

The central atom labelled "biological terrorism" can be unfolded to show its internal structure. We can clearly perceive three sub-graphs of clusters with some connections between them (fig. 3.3).

The topmost part reflects research on the psychological aftermath of september 11, 2001 attacks, namely posttraumatic stress disorders (PTSD). The middle part of the central atom corresponds roughly to two major clusters on "body injuries in terrorist bombing" and "health-care". The lower part of the graph reflects research on potential terrorists attacks using biological and nuclear weapons. The structure of these three

[5]Detailed views of all atoms and network can found on http://demo.termwatch.es

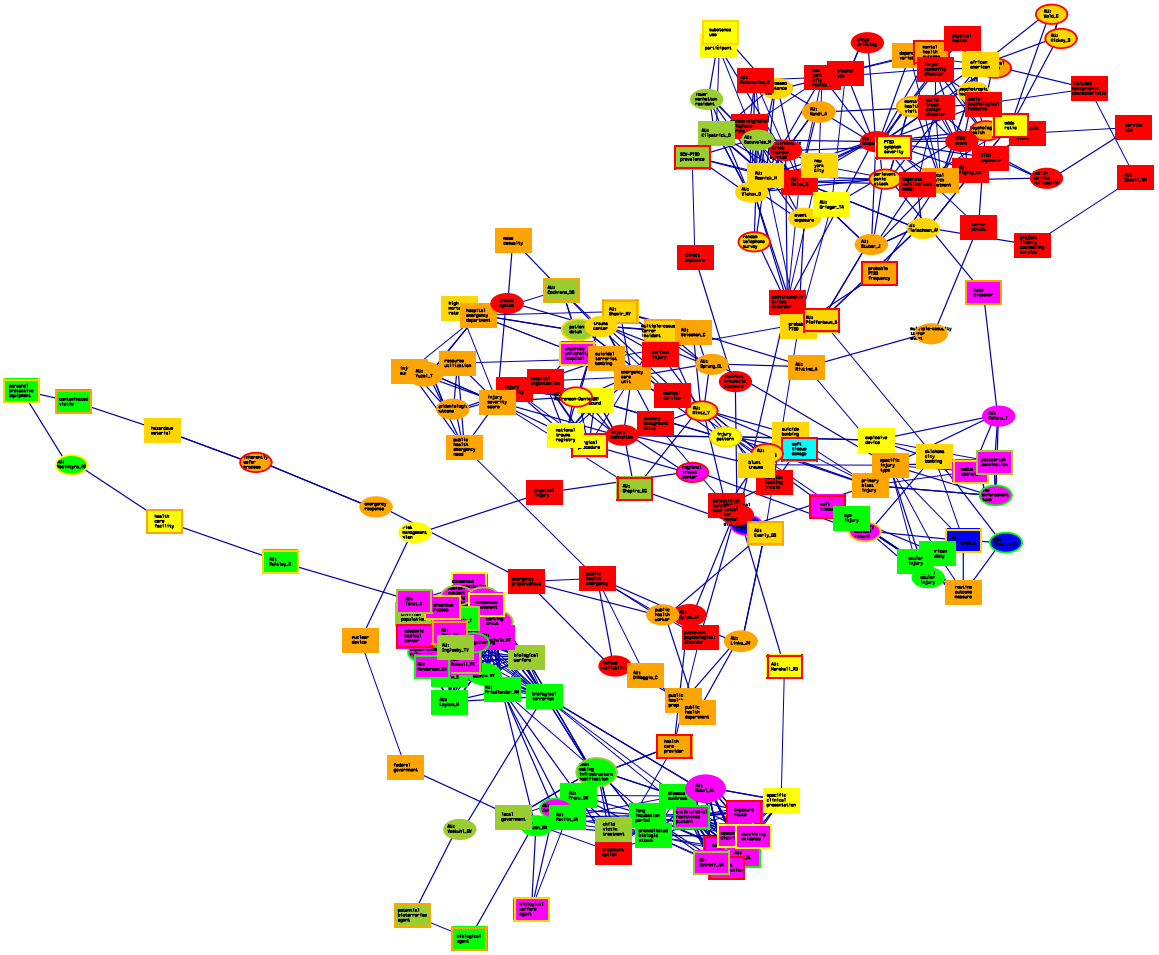

Figure 3.3: Internal structure of the central atom on “biological terrorism”.

sub-graphs echoes to a certain degree the network found in Chen (2006) for the period 1990-2003. Mapping a hybrid network of cited documents and citing terms, he found three major groups of clusters reflecting three research threads: a first thread on “body injuries in terrorist bombing”, a second bigger thread on “health care response to the threat of biological and chemical weapons”, a third biggest and more recent thread on ”pyschological and pyschiatric impacts of the september 11, 2001 terrorist attack” with terms like “United States” and “posttraumatic stress disorder” (PTSD) being very prominent. Globally, these three big threads of research are still present in 2006, albeit with significant changes. Since 2003, the first two threads on “body injuries” and “emergency medical care” have merged into one single thread while a new thread on bio-terrorism has emerged and become more prominent.

The system also computes statistical indicators from the Social Network Analy-

sis Freeman (1977) in order to characterize the relative position of nodes and their importance in the network. Nodes with high betweenness centrality values are possible transitions points from one research thread to another. "posttraumatic stress disorder" (PTSD) is the node with highest betweenness centrality. All other topmost nodes recall major terrorist threats ("traumatic event", "world health", "suicidal terrorist bombing", "biological terrorism", "mass destruction"). The three research threads portrayed by the three sub-graphs in the central atom are present in the first 20 nodes ranked by betweenness centrality: "posttraumatic stress disorder" (1st), "specific injury type" (8th), "primary injury blast" (18th), "biological terrorism" (6th).

The domination of red colour in the upper part of the central atom indicates that the majority of terms in these clusters appeared in the last period (2006). This sub-graph corresponds roughly to the most prominent thread found in Chen (2006) on "September 11" and "posttraumatic stress-disorder" (PTSD). This last term is still very much present three years later as shown by terminological variations found around this term, both in its developed form ("posttraumatic stress disorder symptom") and in abbreviated forms ("probable PTSD frequency", "PTSD symptom severity",“ SCW-PTSD prevalence"). At the center of this sub-graph is the author node "Boscarino JA". His papers focused on psychological effects and PTSD caused by the 9/11, 2001 event. Among the pre-occupying health issues brought to light by this research thread is the increased use of drugs, alcohol and the increase in mental disorder among the population in the area surrounding the World Trade Center. This is evident in the surrounding cluster labels: physical health, psychological health, binge drinking, alcohol use, increased substance use, african-american, posttraumatic stress disorder symptom, psychotropic medication. Boscarino's studies were mostly carried out as sociological surveys by on-line questionnaire administration or telephone surveys (hence a cluster "random digital telephone surveys"). Researchers involved in this topic discovered an increased use of post-disaster medical services to combat PTSD predominantly among the white community, more prone to depression than the non white community. These findings were not yet visible in 2003.

Another difference or evolution observed in our graph and the network proposed by Chen (2006) on the 1990-2003 data is the absence of the cluster "United States". This term has since been replaced by studies focusing on the precise area where the terrorist attack took place, hence the presence of the clusters labelled "new york resident, new york city, lower manhattan resident". It seems that with time, PTSD studies of the 9/11, 2001 terrorist attack have shifted from the nation-level crisis stance (The US was being attacked by terrorists) to a more localised and detailed level - the actual places where the attack took place and its effects on different segments of the population.

3.5.3 Mining closed frequent itemsets on terrorism research

For complexity reasons, it is not possible to extract frequent itemsets whose extension has fewer than three documents, meanwhile we shall see that the atom graph allows us to identify interesting closed itemsets whose extension has only two documents. Using the apriori algorithm in R package, we found 1926 closed itemsets with a support of at least three documents of which 285 have more than three elements (three items). The largest closed frequent itemset without author names is: {*new york city, post-traumatic stress disorder, potential terrorist attack, same traumatic event, world trade center*}. The largest overall has 12 items: *{Parker G,Perl TM,Russell PK, biological terrorism, biological warfare, consensus-based recommendation, emergency management institution, MEDLINE database, nation civilian population, potential biological weapon, working group, world health}*. It appears that both itemsets can be clearly visualized on the central atom.

The graph layout moreover allows us to show how they these frequent itemsets are related and to point out distinct smaller concepts around them. When comparing the central atom structure with closed frequent itemsets, we find out that the upper part of the graph (9/11 & PTSD) is structured around the clique that corresponds to the longest closed frequent itemset without author name. The lower part (bioterrorism) is structurally organized around the clique that represents the longest frequent itemset containing authors items. It also contains the closed frequent itemset *{mass destruction, mass destruction weapon, nuclear weapon proliferation}*. The middle sub-graph focused on "physical injuries and emergency medical care" for victims of terrorist attacks point out formal concepts that connect the two parts of the graph. Apart the frequent item set {*blast lung injury,physical examination,primary blast injury*}, the extension of these formal concepts have only two documents and so cannot be directly computed by R arules library for complexity reasons (memory over stack). However they are essential to understand the relations between the upper and lower part of the graph that are clearly revealed by the visualisation of the graph structure. Finally, all extracted closed frequent itemsets correspond to the cliques in these two sub-graphs of the central atom.

3.6 Conclusion

We have presented a platform for mapping the dynamics of research in specialty fields. The distinctive features of this methodology resides in its clustering algorithm which is based primarily on linguistic (symbolic) relations and on its graph decomposition

algorithm which renders complex terminological graph for comprehensible for domain analysts. The method has been able to identify the most salient topics in two different research domains and uncover the sub-structures formed by persistent and evolving research threads. More importantly, we have shown that it is possible, with limited linguistic resources, to perform a surface analysis of texts and use linguistic relation for clustering. To the best of our knowledge, this represents a unique and innovative approach to text clustering.

The graph decomposition algorithm offers a way of visualizing complex terminological graphs and revealing particular sub-structures contained therein. Mining frequent itemsets, in combination with evaluation by human experts, offer a joint and strong evidence of the significance of the maps produced for the domain.

Chapter 4

Discourse segmentation and recognition of degree of specialization based on rules

4.1 Introduction

This chapter is about collaborating with the linguists of IulaTerm group[1]. Two concepts are considered: discourse segmentation and the degree of text specialization. The related tasks of segmenting a text into units or classifying sentences by degree of specialty can easily be achieved by linguistics with a high degree of agreement among them. The challenge here is to propose a system that simulates a human expert in linguistics, without requiring a heavy learning procedure. One important constraint, is the explain-ability of the results. The system can eventually suggest alternative segments or misclassify sentences if the linguist finds a coherent explanation that makes him revise its reference. The system can eventually suggest alternative segments or misclassify sentences if the linguist finds a coherent explanation that makes him revise its reference.

[1] https://www.upf.edu/web/iulaterm

4.2 Discourse Segmentation

In this section we report our contribution to DiSeg(da Cunha et al., 2012), the first discourse segmenter for Spanish, a collaborative project lead by Iria da Cunha (UNED, Spain). It produces state of the art results while it does not require syntactic analysis but only shallow parsing with a reduced set of linguistic rules. Therefore it can be easily included in applications requiring fast text analysis on the fly. In particular it will be part of the discourse parser for Spanish that we are carrying out. It will be also used in tasks involving human discourse annotation, since it will allow annotators to perform their analysis starting from a unique automatic segmentation. We describe the system, based on shallow parsing and syntactic rules that insert segment boundaries into the sentences. The system performance is evaluated over a corpus of manually annotated texts.

4.2.1 Problematic

The objective is to segment a text into Elementary Dircourse Units (EDUs). We consider them as in Iruskieta et al. (2014), but only those that include at least one verb (that is, they constitute a sentence or a clause). For example, sentence 1a would be separated into two EDUs, while sentence 1b would constitute a single EDU:

1a [The hospital is adequate to adults,]EDU1 [but children can use it as well.]EDU2
1b [The hospital is adequate to adults, as well to children.]EDU1

Furthermore, subject and object clauses are not necessarily considered as EDUs. For example, sentence 2 would be a single EDU:

2 [She indicated that the emergency services of this hospital were very efficient.] EDU1

We have then developed a segmentation tool based on a set of discourse segmentation rules using lexical and syntactic features. These rules are based on: discourse markers, as “while” (mientras que), “although” (aunque) or “that is” (es decir), which usually mark relations of Contrast, Concession and Reformulation, respectively; conjunctions, as, for example, “and” (y) or “but” (pero); adverbs, as “anyway” (de todas maneras); verbal forms, as gerunds, finite verbs, etc.; punctuation marks, as parenthesis or dashes. Finally, we have also annotated manually a corpus of texts to be used as gold standard for evaluation. The elaboration of a gold standard was necessary due to the current lack

of discourse segmenters for Spanish. We thus evaluate DiSeg performance, measuring precision, recall and F-Score over this annotated corpus. We also consider three different baseline systems and a simplified system named DiSeg-base.

4.2.2 Algorithm

Grammar

In this section we present the grammar we implemented on our discourse segmenter, and we show how its implementation has been carried out.

The FreeLing chunker is a bottom-up parser based on Hidden Markov Model (HMM) classifiers, which uses a Context Free Grammar (CFG) and provides shallow parsed trees or chunks with relative depth. The implemented algorithm assigns a priority to each chunk. It determines that the longest and deepest chunks are applied first. The FreeLing chunking grammar includes a set of options to perform extra tasks over parse trees, such as hiding intermediate categories in recursive trees, adding shallowness to relative deep trees, etc. Chunking rules have been designed in order to work with terminal and preterminal nodes, syntactic categories, lemmata and word forms.

Some new rules (41 rules) have been added to the FreeLing chunking grammar in order to recognize some expressions which are considered discourse markers candidates. Discourse markers are classified into two groups: non-ambiguous and ambiguous. Both classes include single word forms (ex.: concretamente "specifically", también "also"), phrases (ex.: en resumen "to summarize", por ejemplo "for example", al contrario "instead of"), multiwords (ex.: a pesar de "in spite of") and composed conjunctions (ex.: así como también "also"). The implementation of all this information in the grammar is carried out by specifying external lexicons or assigning categories to chunks.

disc-mk = RG* <"lista_sadv.txt">. [1]
disc-mk = +SPS00(en), NCFS000(realidad). [2]

The rule [1] has two constraints that the expression analyzed by the chunker must meet. Firstly, this expression must be an adverb (RG). Secondly, this expression must be included in the adverbial expressions lexicon. Example [2] shows a rule used in chunk categorization. This rule defines prepositional phrase parsing between the preposition en ("in") and the noun realidad ("reality") (en realidad "in fact"). As shown in this example, the parenthesis operator "( )" means that words within parenthesis must

be word forms instead of lemmas. Regarding ambiguous discourse markers (ex.: por lo cual "as a consequence"), the grammar only detects them, but they can only be disambiguated in the next task by means of rules taking into account the context.

disc-mk-amb = +SPS00(por), DA0NS0(lo), PR0CN000(que). [3]

In [3], the rule illustrated recognizes the expression por lo que ("as a consequence") as a chunk that may be an ambiguous discourse marker. In other words, in order to solve the categorization of this expression, the context of the sentence needs to be checked.

Implementation

As we have explained, DiSeg implementation relies on Freeling, although we have carried out some modifications into the default grammar of the shallow parser (mainly recategorizations of some elements into discourse markers). Freeling output is then encoded into an XML structure to be processed by perl programs that apply the discourse segmentation rules in a two-step process. First (DiSeg-base), candidate segment boundaries are detected using two simple automata based on the following tags: ger, forma_ger, ger_pas (that is, all possible present participles or gerunds), verb (that is, finite verbal forms), coord (coordinating conjunctions), conj_subord (subordinating conjunctions), disc_mk (recategorizated elements) and grup_sp_inf (infinitives). The only text markers that are used apart from these tags are the period and two words: que ("that") and para ("for"). Second (DiSeg), EDUs are defined using a reverse parsing from right to left where boundaries are considered only if there is a verb in the resulting segments before and after this boundary. Indeed, if all previously inserted boundaries were considered, EDUs without verbs could be generated. Thus, the architecture of the system has several stages: Sentence segmentation (with Freeling) Shallow parsing (with a recategorized Freeling grammar) transformation to xml (with perl programs) segmentation rules application (with perl and twig): detection of segment boundaries / edus definition

DiSeg-1.0 requires FreeLing and it is made of three elements:

1. A grammar for FreeLing.
2. A small perl program to transform FreeLing output into XML.
3. A second perl program that applies the discourse segmentation rules and requires TWIG library for XML.

Since we use very few text marks, our approach should be easily adapted to other Latin languages defined in FreeLing. Moreover, DiSeg-base could be implemented in a CFG, but it would be less computationally efficient. It is only the final reverse parsing that is not CFG definable. In our experiments we have tested to what extend the non CFG module is necessary.

4.2.3 Evaluation

In this subsection we present the gold standard test corpus that we have compiled to perform the evaluation of the system; moreover we show the obtained results, over the medical and terminological sub-corpora.

Gold standard

The gold standard test corpus includes two sub-corpora. The first one consists of 20 human annotated abstracts of medical research articles. These abstracts were extracted from the on-line Gaceta Médica de Bilbao ("Medical Journal of Bilbao"). This sub-corpus includes 169 sentences, 3981 words and 203 EDUs. This sub-corpus was segmented by Iria da Cunha the leader of this project. Another linguist, external to the project, segmented the corpus following the same guidelines. We calculated the precision and recall of this second annotation. Both measures were very high: precision was 98.05 and recall 99.03. Moreover, after short discussions between annotators, a consensus was reached. We use the consensual segmentation as the medical gold standard corpus.

The second sub-corpus includes 10 human annotated abstracts of terminological research articles. These abstracts were extracted from the Proceedings of the Intentional Conference of Terminology in Donostia and Gasteiz in 1972. This sub-corpus includes 125 sentences, 3352 words and 218 EDUs. Once again, this sub-corpus was Iria da Cunha and another linguist (different than the previous one) segmented it following the same guidelines. We calculated the precision and recall of this second annotation. Both measures were very high: precision was 99.014 and recall 99.02. The second annotator had three segmentation mistakes, and, after a short discussion about it, a consensus was reached by both annotators. We use the consensual segmentation as the terminological gold standard corpus.

This gold standard corpus is available at http://diSeg.termwatch.eu.

The statistics of both sub-corpora (sentences, words and EDUs) are similar. Nevertheless, the medical sub-corpus includes 20 documents while the terminological one contains 10. The reason of this difference is that the terminological abstracts are longer than the medical ones, in terms of words.

Experiments with the medical sub-corpus

Firstly, we ran DiSeg over the medical sub-corpus for evaluation and we computed precision, recall and F-Score measures among detected and correct boundaries. Precision is the number of correct boundaries detected by the system over the total number of detected ones. Recall is the same number of correct boundaries detected by the system but divided this time by the total number of real boundaries existing in the gold standard corpus. We did not count sentence boundaries, in order to not inflate the results. For this evaluation, we used three baseline segmenters:

Baseline_0 only considers sentences as EDUs. This is not a trivial baseline since its precision is 100% by definition and four texts in the gold standard have no other type of EDUs.

Baseline_1 inserts discourse boundaries before each coor tag introduced by the Freeling shallow parsing.

Baseline_2 considers both tags indicating coor and conj_subord, but only the last segment at the right of the sentence with a verb is considered as an EDU.

We also consider a simplified system named DiSeg-base, where all candidate boundaries are considered as real EDU ones, even though some generated segments can have no verbs. For Baseline_1, Baseline_2 and DiSeg-base we do not count sentence boundaries.

Our results show that DiSeg full system outperforms DiSeg_base and all the baselines. DiSeg obtains an F-Score of 80% (71% of precision and 98% of recall), while DiSeg_base obtains an F-Score of 74% (70% of precision and 80% of recall). The results obtained by the three baselines are lower: 72% of F-score by Baseline_2 (68% of precision and 82% of recall), 39% of F-score by Baseline_1 (33% of precision and 70% of recall) and 62% of F-score by Baseline_0 (100% of precision and 49% of recall).

F-Score differences are statistically significant according to the pairwise Student test at 0.05 between the two versions of DiSeg and at 0.01 among DiSeg and the three

baselines (Baseline_2, the most sophisticated baseline, obtains the best results). These results are similar to those obtained by the discourse segmenter for English developed by Tofiloski et al. (2009): 93% of precision, 74% of recall and 83% of F-Score.

Although we considered these quantitative results were good, we carried out a qualitative analysis in order to detect the main performance problems. After this qualitative analysis of the results, we developed three more symbolic rules to try to solve the main systematic segmentation errors. The rules are applied in the post-processing stage (EDUs definition), so they can increase results precision. In this way we try to optimize the system, that we call now DiSeg-1.0. We have applied DiSeg-1.0 over the same medical sub-corpus, in order to check if the results improve. Tthe results of DiSeg-1.0 outperform the results of the previous version of the system, DiSeg. The F-score of DiSeg 1.0 is 96% (97% of precision and 96 of recall) and the F-score of DiSeg is 80%, so there is a difference of 16 points.

Experiments with the terminological sub-corpus

Once we have tested the system with the medical sub-corpus, we have decided to apply it over another corpus including documents from a very different domain, the terminological one. Both sub-corpora are specialized, but they correspond to very different domains: a technical or scientific domain vs. a humanistic or linguistic domain. The reason of this selection was that we wanted to check that DiSeg-1.0 is suitable to segment any kind of text. We have applied DiSeg-1.0 over this terminological sub-corpus. We have used a baseline to carry out this evaluation. This Baseline_0 was performed in the same way we have done in the previous experiment with the medical sub-corpus. In this experiment, DiSeg-1.0 obtains an F-score of 91% (95% of precision and 87% of recall) and Baseline_0 obtains an F-score of 62% (100% of precision and 49% of recall). Thus, the results obtained by our system are very high (much more than the baseline), and this means that our system can be used to segment texts of different domains.

4.3 sentence specialization level detection

4.3.1 Problematic

Nowadays, compilation of Languages for Specific Purposes (LSP) corpora, that is, corpora including specialized texts, is necessary to carry out several tasks, such as: ter-

minology extraction, developing of specialized dictionaries or lexicons, preparation of ontologies, etc. This corpora compilation means human participation, since, until now, professionals or specialists were the responsible ones to decide if the documents from their domains were specialized or non-specialized. This situation entails some negative aspects, mainly the necessary time and human effort.

In this line, search engines users often need to find specialized documents spending lots of time searching this type of documents manually. Some efforts have been done to solve this problem. For example, Google Scholar allows us to obtain academic documents, as thesis, research papers, abstracts, etc. Nevertheless, the system is not based on the content of the documents, but on some external aspects as the format (the obtained texts are mainly pdf documents) and some indexing research aspects (since they work with editors publishing academic material and with academic and scientific publications).

But what is a specialized text? Cabré (2002) mentions some variables that have to be considered in order to answer this question: the text author, the potential reader, the structural organization and the lexical units' selection. She affirms as well that there are two types of variation of the specialized texts: horizontal variation (determined by the subject) and vertical variation (determined by the specialization level). With regard to the second one, three specialization levels can be considered: high (specialized writer and specialized receiver), medium (specialized writer and semi-specialized receiver, that is, for example, students) and low (specialized writer and non-specialized receiver, that is, general public).

It is important to note that, for a text to be considered as specialized (with a high, medium or low level), the writer (or speaker) of that text has to be a specialist of the domain, since specialists are the only ones who have the necessary deep knowledge to write specialized texts. Articles in newspapers may deal with technical subjects, as, for example, economics, medicine or law. However, if they are written by journalists, they cannot be considered as specialized, because general journalists don't use to have the "conceptual and lexical control" of these domains.

There are several theoretical works about differences between general and specialized texts. Most of them consider that lexicon is the most distinguishing factor (besides being the most visible) to carry out this differentiation. It is well-known that terms (units of the lexicon with a precise meaning in a particular domain Cabré (1999)) show the specialized content of a subject; therefore, they appear inevitably in texts of their domain. Thus, other characteristic features of specialized texts (as grammatical features, both morphological and syntactic) can be considered as specific of these texts.

Features as verbal flexion related to grammatical person, verbal tense or verbal mode have been underlined in some works Kocourek (1991). Some authors, using small corpora, have established some grammatical phenomena that may differentiate specialized texts. In some cases, they have considered only a very limited number of features of a single category; in other cases, a scarce number of texts has been analyzed manually. Hoffmann (1976) analyzes the frequency of names and verbs into a general corpus and a specialized corpus. Some authors have studied verbs into specialized French corpora Coulon (1972); Cajolet-Laganière and Maillet (1995); L'Homme (1993, 1995). The works of Cabré et al. (2010); Cabré (2005) are the first ones where this subject is studied using a bigger corpus (two millions of words). They conclude that certain grammatical features, besides lexicon, have a strong potential to differentiate specialized texts from non-specialized texts.

As mentioned above, there are some theoretical studies about the characterization of specialized and non-specialized texts. Nevertheless, at our knowledge, there are not works about the automatic differentiation of both types of texts. The aim of this work is to develop the first tool for automatic specialized vs. non-specialized texts differentiation, based on the content of documents. To develop this tool, firstly we have compiled a corpus, including Spanish specialized and non-specialized texts in economics. Secondly, we have splitted both corpora into two sections: training and test. Finally, we have used machine learning techniques to develop two different strategies to automatically differenciate between specialized/non-specialized texts with association rules combining grammatical and lexical features. Our results show that both strategies are suitable to differentiate specialized vs. non-specialized texts, although, as we will show, the type of corpus influences the results. We consider that the automatic tool we have developed will be very useful for the two tasks mentioned at the beginning: the automatic constitution of specialized corpora and the optimization of specialized search engines.

4.3.2 Methodology

The corpus was divided as follows:

1. A sub-corpus including texts from the specialized domain of economics, mainly scientific papers, books, theses, etc. (with 292,804 tokens corresponding to 9,243 sentences).
2. A sub-corpus with non-specialized texts from the economics subsection of Spanish newspapers (with 1,232,512 tokens corresponding to 36,236 sentences).

These texts have been extracted from the Technical Corpus of the Institute for Applied Linguistics[2] (IULA-CT) of the Universitat Pompeu Fabra of Barcelona. It consists of documents in Catalan, Spanish, English, German and French, although the search through bwanaNet is at the moment restricted to the first three of these languages. It contains texts of several specialized domains (economics, law, computing, medicine, genome and environment) and plain texts from newspapers. All the texts are tagged with POS tags. This corpus is accessible on-line via http://bwananet.iula.upf.edu/. Further details on these resources are shown at Vivaldi (2009).

All the texts were tagged with POS tags.

We then have selected some linguistic features that may be characteristic of specialized texts and non-specialized texts. We have used the features detected by Cabré et al. (2010) and Cabré (2005). Table 4.1 shows them. The full meaning of these POS tags can be seen on the following URL: http://www.iula.upf.edu/corpus/etqfrmes.htm.

Some POS tags are produced by subespecification of the full tag (ex. "A" is a subespecification of "AMS", "AMP", etc.). The machine learning approach that we have used is based on association rules, one of the most-known methods to detect relations among variables into large symbolic (i.e. non numerical) data Amir et al. (2005).

We choose to work on sentences instead of entire documents. Indeed, documents can be classified using contextual information about their structure or statistical information about their specific vocabulary. At sentence level, none of these informations can be used. Therefore, the application that we propose not only allows to classify texts, it also allows us to look for technical/non-technical statements inside any document type.

In the third place, we have evaluated the results. This evaluation is based on the capacity of the tool to differentiate sentences coming from specialized texts from others over the mentioned test corpora (specialized and non-specialized).

4.3.3 Experiments, Settings and Results

In our machine learning experiments with association rules and n-grams method, we have randomly selected 9,000 sentences from each corpus (specialized and non-specialized). Therefore the experiment has been carried out on a set of 18,000 sentences with a total

[2] http://www.iula.upf.edu

Table 4.1: Linguistic features used in our work.

POS	Tag meaning
A	Determiner
C	Conjunction
D	Adverb
E	Especifier
JQ	Qualifier adjective
J	Adjective
N4	Proper noun
N5	Common noun
P	Preposition
R	Pronoun
T	Date
VC	Verb (participle)
V1P	Verb (first person, plural)
V1S	Verb (first person, singular)
V2	Verb (second person)
V	Verb
X	Number

of 112,870 tokens. We have used the 90% of both corpora for training and the 10% for test, replying this split 30 times at random.

For the training with association rules, we have used sentences level (although we have tested that only sentences with more than six words can be classified). We have employed a machine learning strategy based on the combination of lexical features (lemmas) and grammatical features (POS tags).

Table 4.2 shows an example of plain text and its corresponding generated test corpus text. In bold we have marked the category GEN, which is indicating that this sentence is classified as part of a non-specialized text. Observe that "Plain text" section includes the sentence as found in the general corpus while the "Attributes generated from text" section includes just a list of the lemmas/tags found in such sentence.

We consider association rules of the form X $\Rightarrow$ D, where X is a set of at most 5 lemmas and/or tags, D is the decision: SPE for specialized and GEN for general. For a rule to be valid, X has to be included in more than 0.5% of the sentences (this is called the support of the rule) and more than 90% of these sentences that include X have to be in category D (this is called the confidence of the rule). Since the right part

Table 4.2: Example of economic plain text and attributes generated from text.

Plain text
Tras el acuerdo con los pilotos, la dirección de Alitalia concluyó ayer de madrugada la negociación con los sindicatos del personal de tierra, que aceptaron 2.500 despidos (la propuesta inicial era de 3.500), la congelación de los salarios durante dos años y el bloqueo del fondo de previsión social durante el mismo periodo, para evitar la quiebra de la compañía.
Attributes generated from text
GEN ser congelación despido previsión tierra dos dirección el tras para quiebra periodo negociación mismo piloto bloqueo = salario A Alitalia C D de N4 N5 personal compañía fondo P R que JQ V propuesta num X social con ayer aceptar madrugada sindicato concluir año inicial durante acuerdo y evitar

of the rule is restricted to a few numbers of categories, we shall refer to these rules as decision rules. This kind of rules can be computed using standard GPL packages like "Apriori" by Christian Borgelt (http://www.borgelt.net/apriori.html). Our experiments over the economic corpus show that this strategy allows us to obtain 46,148 decision rules. It appears that:

- 60% of the rules induce category SPE, which means that there are more implicit decision rules among specialized texts than non specialized ones.
- 78% of the rules include at least one grammatical tag which shows that this information is significant to distinguish between these two categories.

Here is a sample set of 10 rules randomly extracted from the total list of decision rules for the economic corpus. Rules are given in Prolog format: the decision is on the left and the two figures give respectively the support and the confidence of the rule.

SPE ← europea N4 JQ N5 (50, 100.0)

SPE ← millones X JQ P (70, 100.0)

GEN ← anunciar N4 P = (80, 98.3)

GEN ← ayer uno R N4 (10, 100.0)

SPE ← función C JQ D (12, 93.1)

GEN ← Gobierno haber VC V (60, 100.0)

GEN ← España que P = (100, 100.0)

SPE ← embargo sin de N5 (70, 100.0)

SPE ← internacional a R N5 (12, 90.8)

GEN ← presidente en R JQ (80, 93.0)

Therefore each rule indicates that if a given set of lemmas and tags is included in one sentence, there is a specific probability to classify the sentences as general (GEN) or specialized (SPE). As an example, the first rule may be read as follows: if the sentence under analysis includes the lemma "europea" and words with the POS tags "N4", "JQ" and "N5", then such sentence may be classified as specialized (SPE). The coverage of this rule is 50% with a 100% of precision.

Once this set of rules is available, it is possible to build a classifier that, given a sentence, looks for the set of rules that match the sentence and chooses the rule that has the highest confidence. One important feature of this type of classifier is that it indicates when it cannot take a decision. Finally, for a given text under analysis, if more than a half of the sentences it contains belong to a given category the text is considered to belong to such category.

As a variant of this basic classifier (Classifier 1) we have developed a variant that only takes into accout those rules including at least one POS tag (Classifier 2). In this way it is possible to evaluate the actual impact of using POS tags as a classifier atribute.

Results

Results are shown in Table 4.3.

Table 4.3: Results of Classifier 1 over the economics corpus.

	Precision	Recall	F-Score
GEN	0.7602	0.8671	0.8137
SPE	0.8875	0.7239	0.8057
Average	0.8190	0.7890	**0.8040**

We have carried out another experiment over the economics corpus, using for the classifier (Classifier 2) only the association rules including at least one grammatical feature (POS tag). This is a subset of 36,217 rules (78%). Results obtained by Clasifier 2 over the economics corpus are shown in Table 4.4.

Table 4.4: Results of Classifier 2 over the economics corpus.

	Precision	Recall	F-Score
GEN	0.7582	0.8959	0.8213
SPE	0.8749	0.7182	0.7889
Average	0.8166	0.8071	**0.8051**

This evaluation indicates that elimination of rules exclusively based on lemmas does not significantly degrade classifier performance. In fact, is seems that it lightly improves the average F-score (from 0.8040 to 0.8051).

Obtained results with both strategies are good over the economics corpus, although results with n-grams distances are a bit better than using association rules (0.8051 vs. 0.8385). Nevertheless, the association strategy has one advantage: the generated rules are humanly understandable and interpretable. The n-grams strategy offers only n-grams of characters, that is, unintelligible textual short passages.

Table 4.5: Results of n-grams of string classifier over the sexuality corpus.

	400K 13-grams			500K 15-grams		
	Precision	Recall	F-Score	Precision	Recall	F-Score
GEN	0.7999	0.8121	0.8058	0.8102	0.8156	0.8128
SPE	0.8370	0.8257	0.8312	0.8412	0.8361	0.8385
Average	0.8184	0.8189	0.8185	0.8257	0.8258	**0.8257**

4.4 Conclusion

This chapter gives an example of interdisciplinary work between computer science and linguistics. While this was common in the 80's, the usage of machine learning black boxes relying on big data resources made these collaborations less frequent. As a result computer scientists focus on improving scores and linguists focus on tagging large resources. However these are two cases where state of the art classifiers are difficult to apply. This is due to the lack of resources and the span of the contextual windows required to solve these problems.

Meanwhile, collaborative work leads to results on both disciplines. New algorithms are created to analyze instantly large volumes of texts. Linguistics get instant feedback and discover new cases that feed their research.

Part II

Focused retrieval

Chapter 5

Interactive Query reformulation

5.1 Introduction

Most of the scholar evaluation campaigns in IR focus on fully automatic approaches with automatic evaluations. This can lead to artificial problems since in most real user cases, some interaction can be expected with the user. In this chapter we will show how a minimalist interaction can reveal users' intention and unexpectedly improve the results. This opens new perspectives for recent IR commercial engines, like DuckDuckGo or Qwant, that focus providing full privacy but rely more on query analysis to understand the user's intention.

The issue of how to represent queries and documents has been a recurrent one in information retrieval (IR). System designers usually have the choice between representing queries and documents as single units (bag-of-words) or with longer patterns which can be noun phrases, multiword terms, n-grams, fixed expressions, collocations or text spans. The choice of longer text units naturally raises the question of how to first identify them from queries and subsequently from documents, and how to represent them within an IR model. Two main approaches have been explored to this end: the linguistic model Perez-Carballo and Strzalkowski (2000) that in turn raises the issue of the role of Natural Language Processing (NLP) in IR; and statistical or probabilistic Language Models (LM). The LM approach Metzler and Croft (2003) was inspired by the most successful approaches issuing from research in speech recognition.

Previous experiments carried out within the framework of TREC Voorhees (1999); Sparck-Jones (1999); Perez-Carballo and Strzalkowski (2000) tended to conclude that retrieval performance has not been enhanced by adding NLP, especially syntactic level

of processing. The problem lies in determining the level of NLP needed, on which text units to implement it, whether to implement NLP on both queries and documents and at what stage (whole collection or only on an initial set of returned documents). Previous research also concluded that a deep syntactic representation of queries and documents is not useful to achieve a state-of-the-art performance in IR Smeaton (1999). It may on the contrary degrade results. On the other hand, performance can be boosted by better representing queries and documents with longer phrases using shallow NLP. In some cases, even a well-tuned n-gram approach can approximate the extraction of phrases and may suffice to boost retrieval performance.

Up until 2004, the dominant model in IR remained the bag-of-words representation of documents which continued to show superior performances in IR. However, a series of experiments carried out on several document collections over the past years are beginning to show a different picture. Notwithstanding the apparent success of the bag-of-word representation in some IR tasks, it is becoming clear that certain factors related mostly to query length and document genre (general vs technical) influence the performance of IR systems. For instance, Perez-Carballo and Strzalkowski (2000); Mishne and de Rijke (2006) showed that representing queries and document by longer phrases can improve systems' performances since these text units are inherently more precise and will better disambiguate the information need expressed in the queries than lone words.

Furthermore, Perez-Carballo and Strzalkowski (2000) concluded that the issue of whether or not to use NLP and longer phrases would yield better results if focused on query representation rather than on the documents themselves because no matter how rich and elaborate the document representation, a poor representation of the information need (short queries of 1-2 words) will ultimately lead to poor retrieval performance.

Based on these earlier findings, we wish to investigate the issue of representing queries with a particular type of phrase which are Multiword Terms (MWTs). MWTs is understood here in the sense defined in computational terminology Kageura (2002); Castellvi et al. (2001) as textual denominations of concepts and objects in a specialized field. Terms are linguistic units (words or phrases) which taken out of context, refer to existing concepts or objects of a given field. As such, they come from a specialized terminology or vocabulary Ibekwe-SanJuan (2006). MWTs are thus terms of length >1. MWTs, alongside noun phrases, have the potential of disambiguating the meaning of the query terms out of context better than single word terms or statistically-derived n-grams and text spans. In this sense, MWTs cannot be reduced to words or word sequences that are not linguistically and terminologically grounded. An initial selection of MWTs from queries is used in an Interactive Query Expansion (IQE) process to acquire more

MWTs from top n-ranked documents. The expanded set is submitted to standard IR Language Models for document ranking. Our approach is tested on two corpora: the TREC Enterprise track 2007 and 2008 collections, and INEX 2008 Ad-hoc track. We chose as baseline against which to compare our IQE approach, an IR engine based on the language model using Dirichlet smoothing. The Indri IR system Metzler et al. (2005) in its default mode applies this language model. Indri was also used as baseline in TREC terabyte[1]. The idea was to test our IQE approach against a strong baseline that competes favorably with the best systems in current IR evaluation campaigns. The results obtained on the Wikipedia corpus in the INEX Ad-hoc track are particulary promising.

The rest of the paper is structured as follows. Section §5.2 offers a synthesis of earlier studies on the effectiveness of phrase and query expansion (QE) in IR. Section §5.3 presents our language model and its application to the IR tasks. Section §5.4 describes the application of our IR model to the TREC Enterprise track 2007 and 2008 collections for document search task. Section §5.5 presents the focused retrieval tasks on the Wikipedia collection in the INEX 2008 Ad-hoc track. Finally, section §5.9 discusses lessons learned from these experiments.

5.2 Related work

Since 2004, new results in IR changed the general opinion on the effectiveness of phrase-based query representation but not on the usefulness of syntactic analysis for standard document retrieval. We synthesize here previous studies that experimented the combined use of query representation by phrases (defined loosely here as any thing other than lone words) with QE. Also of particular interest to us are studies that employed language models (LM) with smoothing mechanisms.

5.2.1 Effectiveness of query representation by phrases

It has been show in Mishne and de Rijke (2006) that Web retrieval of HTML elements based on short focused queries can be boosted by considering sub-phrases of the query. The experiment was carried out using a standard vector model, phrases being considered as special single features. For efficiency reasons, the authors estimated the $idf(t)$ of the phrases $t = (w_1, ..., w_n)$, as the minimum $\min_i idf(w_i)$ of their components. The

[1] `http://stefan.buettcher.org/trec-tb/`

model works on the assumption of statistical independence between features, no NLP was involved. By combining standard language models (LM) Metzler and Croft (2003) with inference networks of the InQuery IR engine Callan et al. (1992, 1995), Metzler et al. (2005); Metzler and Croft (2005) showed that phrase-based queries performed effectively on large-scale collections such as the Web. Indeed, inference networks allows the expression of more complex term dependencies in the query and avoids the assumption of term independence. The LM used in Metzler and Croft (2003) relies on the usual multinomial distribution model but uses Jelinek-Mercer smoothing as defined in Zhai and Lafferty (2004). It appeared later that Dirichlet smoothing, also studied in Zhai and Lafferty (2004) has a better theoretical and formal background. However, in both types of smoothing, it was necessary to choose different parameters according to the phrase length. Experiments conducted in Metzler and Croft (2003); Metzler et al. (2005) and then in Eguchi and Croft (2009) show that phrase-based structured queries are able to filter out most of the noisy documents when the collection is large enough to estimate their likelihood. Hence, phrase-based queries appear to be efficient on large but noisy collections. Still these phrases were selected based on a probabilistic language models without requiring any NLP. In Metzler and Croft (2007), the same model was used to expand queries by automatically finding related phrases in top ranked documents by the initial query. Phrases up to three words were considered but it appeared that it was sufficient to consider independent single terms for query expansion (QE). Hence, according to this experiment, phrases are efficient to better express real user's queries but they appeared unnecessary for QE. The above two approaches Mishne and de Rijke (2006); Metzler et al. (2005) did not make use of advanced NLP techniques like the ones employed in Perez-Carballo and Strzalkowski (2000), but relied on vector calculus and probabilistic models. In this context, phrases can be any sequence of words. However, most of the phrases considered in Metzler and Croft (2007) appear to be well formed noun phrases (NPs) among which some corresponded to MWTs. In both approaches, Mishne and de Rijke (2006); Metzler et al. (2005), it was necessary to adjust model parameters to phrase length (in the idf or smoothing components).

On the other hand, Perez-Carballo and Strzalkowski (2000) showed that the use of more advanced NLP techniques coupled with IQE can boost retrieval performance. They explored the effectiveness of indexing documents' summaries selected by users in an IQE process, then re-indexed the summaries using NLP techniques for query representation. Previously, Strzalkowski et al. (1999) had developed an Interactive Query Expansion (IQE) system that was ranked among the eight best manual systems in the ad-hoc track of TREC-8 conference. In this study, user interaction was limited to 10 minutes. Within this time frame, a set of 30 abstracts extracted from the top ranked documents from the initial query were presented to the user who had to remove those abstracts that were not relevant. The remaining abstracts were added to the

query and the resulting text processed with the same NLP tools that was used to index the documents. The expanded query was then re-submitted. According to the authors, this simple and short interaction was sufficient to improve dramatically the performance of their linguistically-based IR system. However, the authors also mention that the expanded queries also significantly improved the performance of more statistical systems on TREC 6, 7 and 8 data. The authors left open the question as to whether the improvements were due only to the user's selection, to the automatic summarizer or to the linguistic indexing of documents.

Vechtomova (2005) applied NLP in order to extract noun phrases (NPs) used in an IQE process. The IQE approach described in her study shares similar points with that of Perez-Carballo and Strzalkowski (2000) except that instead of using the abstracts of the top n-ranked documents to expand the queries, Vechtomova (2005) extracted NPs from query topics using a part-of-speech tagger and a chunker. She tested different term weighting functions for selecting the NPs: idf, C-value and log-likelihood. We refer the reader to Knoth et al. (2009) for a detailed description and comparison of these measures. The ranked lists of NPs were displayed to the users who selected the ones that best described the information need expressed in the topics. Documents were then ranked based on the expanded query and on the OKAPI probabilistic model Jones et al. (2000). By setting optimal parameters, the IQE experiment in Vechtomova (2005) showed significant precision gains but surprisingly only from high recall levels.

5.2.2 Cognitive biases in IQE experiments

Much research effort has been expended on ways to assist users in formulating queries by means of low charge cognitive operations Marchionini (1992), especially in the case of vague information needs. As widely observed, users often prefer to start their search with an imprecise or vague query, browse the top ranked documents and eventually reformulate the query. The easiest way to reformulate a query is then to add terms found in relevant top n documents. This can be done automatically (AQE) or interactively (IQE). However, IQE in itself is not a trivial process. For it to be effective and be able to compete with wholly automated IR procedures, some precautions need to be taken. For instance, Ruthven (2003) observed that human experts were not necessarily better placed to select and weigh good candidate terms for QE. In particular, he underlined *"how difficult it is to select a set of expansion terms that will perform better than AQE or no query expansion"* and concluded that *"simple term presentation interfaces are not sufficient in providing sufficient support and context to allow good query expansion decisions. Interfaces must support the identification of relationships between relevant material"*.

These observations are not surprising if we think of the task of formulating the best query as a case of "judgement under uncertainty" as defined in Tversky and Kahneman (1990); Kahneman et al. (1981). When humans are presented with questions of the type "what is the probability that object A belongs to class B" (in our case B is the set of all terms in relevant documents), they tend to apply several cognitive heuristics like *representativeness* of A ("Does A have some conceptual similarity with concepts in B?") and *availability* ("can I remember or easily find relevant documents containing instances of A ?"). However, these two heuristics lead to biased decisions. In particular they tend to overestimate highly frequent or abstract events A. It has been shown in Tversky and Kahneman (1990) that in the stress of real situations, humans tend to ignore basic numerical facts like conditional probabilities B/A even when they are aware of it. They consider them only when no other information is available about event A. But if they have some knowledge about A, even when this knowledge is totally uninformative about the relation between A and B, humans will base their decision on this knowledge with little or no regard for the prior probabilities of the categories. Cognitivists also observed that humans tend to favour abstract terms over concrete one because of the availability heuristics. As pointed out in Tversky and Kahneman (1990): "*It seems easier to think of context in which an abstract concept is mentioned (love in love stories) than to think of contexts in which a concrete word (such as door) is mentioned.*" These cognitive biases imply that IQE will not necessarily be more effective than AQE since automatic procedures can easily avoid these two biases by computing standard $tf.idf$ or conditional probabilities. These cognitive observations can explain why users in IQE tend to ignore term statistics displayed by the QE interface. In Vechtomova (2005), it was observed that ordering phrases based on *idf* average, C-value or likelihood measures had no impact in the way users selected them. It has also been shown that users did not need to see the context from which phrases were extracted, thus suggesting that they based their decision of selecting a phrase only on its meaning or at least on their own conception of its meaning. This would suggest that the choice of meaningful phrases or of MWTs lead to self-explanatory text units that should support good IQE strategies regardless of context or of a particular term weighting function. Our aim is to show that a careful selection of the particular type of text units with which to represent queries, here MWTs, can overcome most of the difficulties normally encountered in an IQE process.

5.3 Combining Automatic and Interactive Query Expansion

5.3.1 Motivations for our study

Multiword terms (MWTs) are a particular phrase type that designate domain objects and concepts. As such, they have much lower frequency occurrence compared to that of statistically-derived phrases or single word terms. This low frequency property should help avoid the representativeness bias Tversky and Kahneman (1990); Kahneman et al. (1981). Secondly, their high specificity will also help avoid the second bias induced by the high cognitive availability of common or abstracts terms. Therefore, if the user focuses on what s/he thinks are real MWTs, he should avoid the two cognitive biases and will select terms that an automatic QE procedure will normally reject because of its lack of semantic knowledge and their low frequency. On a linguistic level, terms are a subgroup of phrases (mostly noun phrases) that correspond to domain concepts. Hence, it is not always possible to distinguish them from ordinary noun phrases based solely on their morpho-syntactic composition. Although many term extraction tools exist, what they extract are candidate terms. It is often left to humans to distinguish real domain terms from general noun phrases. At this exploratory phase of our study on the effectiveness of MWTs and IQE for IR, we opted for a human selection of MWTs. If our methodology proves effective, we have tools and means of automating the MWTs selection phase in the future (see 5.6). In the sophisticated automatic query expansion (AQE) system introduced in Metzler and Croft (2007), the use of multi word phrases in the expansion procedure did not significantly improve the system performance but it did not damage it either. Hence, we can expect that the action of a user that keeps on adding MWTs to the initial query and then let the probabilistic IR engine evaluate their likelihood in documents should at least not handicap its performance. We want to show here that it can significantly improve it. Based on the results from previous studies, we emit three hypotheses about the effectiveness of MWTs for QE. They should:

1. be efficient in focused retrieval tasks according to Mishne and de Rijke (2006),
2. allow a better handling of noise in large corpora following Metzler and Croft (2005),
3. be adapted to incremental IQE (where the user keeps on adding more terms to the initial query) following Vechtomova (2005).

We also expect IQE based on MWTs to be sufficiently robust to avoid multiple

parameter tuning. According to Metzler and Croft (2005); Mishne and de Rijke (2006), hypotheses 1 and 2 are valid for phrases. In this chapter, we investigate if they are also valid when we restrict phrases to well formed MWTs. If the previous assertion is true, we also want to evaluate the maximal gains in precision that we can expect in an IQE process based exclusively on MWTs. For that, we will consider different phrase structures for query representation, from plain lists of MWTs to more complex inference networks. We also hypothesize that an IQE procedure focused on MWTs should furthermore avoid the difficulties listed in 5.2.2: the cognitive biases described in Tversky and Kahneman (1990); the selection problem described in Ruthven (2003) and not be reliant on contextual information as observed in Vechtomova (2005). In the sequel, we describe our methodology that combines IQE based on MWTs with Automatic Query Expansion (AQE) since the two are complementary. IQE allows us to apply users' semantic knowledge to the expansion procedure but it may also introduce human-biases. AQE on the other hand avoids these biases but is prone to ignoring obvious semantic relations among MWTs.

5.3.2 Language Model

Language models are widely used in NLP and IR applications Ponte and Croft (1998); Jones et al. (2000). In the case of IR, smoothing methods play a fundamental role Zhai and Lafferty (2004). We shall first describe the probability model that we use.

Document Representation: probabilistic space and smoothing

Let us consider a finite collection $\mathcal{D}$ of documents, each document D being considered as a sequence $(D_1, ..., D_{|D|})$ of $|D|$ terms D_i from a language $\mathcal{L}$, i.e. $\mathcal{D}$ is an element of $\mathcal{L}^\star$, the set of all finite sequences of elements in $\mathcal{L}$. Our formal framework is the following probabilistic space $(\Omega, \wp(\Omega), P)$ where Ω is the set of all occurrences of terms from $\mathcal{L}$ in some document $D \in \mathcal{D}$ and P is the uniform distribution over Ω:

$$\Omega = \{D_i : D \in \mathcal{D}, 1 \leq i \leq |D|\} \tag{5.1}$$

$$(\forall A \subseteq \Omega) P(A) = \frac{|A|}{|\Omega|} \tag{5.2}$$

LMs for IR rely on the estimation of the a priori probability $P_D(q)$ of finding a term $q \in \mathcal{L}$ in a document $D \in \mathcal{D}$. Indeed in LM, each document infers a different probability distribution. There are two trivial ways of defining such distributions. The first one is

to set $P_D(q) = P(q|D)$, therefore we have:

$$P_D(q) \quad = \quad \frac{f_{q,D}}{|D|} \tag{5.3}$$

where $f_{q,D} = |i : D_i = t, 0 < i \leq |D||$ is the frequency of q in D. The drawback here is that for any $q \notin D$, $P(q|D) = 0$. This would lead to IR systems in which only documents containing all terms in the query would be retrieved and where the absolute frequency of term t in the document collection $\mathcal{D}$ would not be taken into account. Obviously, a kind of "inverse document frequency" (*idf*) component is missing in this trivial LM.

An opposite trivial way to define $P_D(q)$ is to set it to $P(q)$:

$$P_D(q) \quad = \quad \frac{f_{q,.}}{|\Omega|} \tag{5.4}$$

where $f_{q,.} = \sum_{D \in \mathcal{D}} f_{q,D}$. This way $P(q|D) > 0$ for any term q occurring at least once in some document, but we would have $P(q|D) = P(q|D')$ for any pair D, D' of documents. Therefore such a radical model is not an appropriate framework to define IR ranking functions. To build efficient document ranking function based on LM, it is necessary to define $P_D(q)$ by combining both equations 5.3 and 5.4. This is called smoothing Zhai and Lafferty (2004) and there are as many ways of doing it as there exists different variants of $tf.idf$-like formulae in IR vector space model Baeza-Yates and Ribiero-Neto (1999). Each variant requires the estimation of several parameters depending on document collection characteristics. In Metzler and Croft (2003), it is the Jelinek-Mercer smoothing that is used. It is a mixture of previous trivial probabilities distributions (5.3) and (5.4) defined as follows:

$$P_D(q) \quad = \quad (1 - \lambda)P(q|D) + \lambda P(q) \tag{5.5}$$

λ_D being some constant in $[0, 1]$.

But in Metzler et al. (2004), the Dirichlet smoothing method is preferred because it can be viewed as a maximum *a priori* (MAP) document probability distribution while considering a multinomial model or a multiple bernoulli model. It is defined as follows:

$$P_D(q) \quad = \quad \frac{f_{q,D} + \mu \times P(q)}{|D| + \mu} \tag{5.6}$$

where μ is an integer.

In fact, equation (5.6) shows that Dirichlet smoothing consists in randomly expanding document D with a sample E_μ of μ terms outside D in order to take into account

the probability of finding q outside D. Indeed, $\mu \times P(q)$ gives an estimation of the frequency of q among the set of μ terms and we have: $P_D(q) \sim P(D \cup E_\mu)$ for $\mu > 1000$.

This suggest that Dirichlet smoothing should be robust in QE procedures since the μ parameter acts as a sample size to estimate term frequency among the whole population. It also indicates that μ should not be much greater than $|D|$ because we would have $P_D(q) \sim P(q)$ like in equation (5.4), and not too small because we would have $P_D(q) \sim P(q|D)$ like in equation (5.3). Another important feature of Dirichlet smoothing is that it takes into account document size. Indeed, in a mixture approach like in (5.5), the contribution of each term is fixed by λ parameter and completely relies on it. λ should then be adapted for each document. Meanwhile, in Dirichlet smoothing, P_D is likewise a distribution of probabilities over $D \cup E_\mu \subseteq \Omega$ whose size varies with D since $|E|$ is fixed. Therefore, Dirichlet smoothing favors short documents, the estimation of $P_D(q)$ involving a denominator of the form $|D \cup E|$. So it should be adapted to focused retrieval where the system has to point out short passages or document elements answering the query.

Both LMs and corresponding smoothing take into account the overall term frequency like the *idf* component in vector model and the document length. However, Zhai and Lafferty (2004) have observed that Dirichlet smoothing performs better on short queries while Jelinek-Mercer smoothing seems to better handle verbose queries. This reinforces the intuition that Dirichlet smoothing is adapted to estimate the probability distributions induced by a document, while the mixture approach is better for query modeling. As expected, it has also been observed that on concise queries, the performance curve tails of Dirichlet smoothing are stable, so it seems less sensitive to μ parameter choice than Jelinek-Mercer smoothing is for λ choice. Based on these empirical results, Zhai and Lafferty (2004) have introduced a two stage smoothing that combines both approaches, however for an IQE procedure based on MWTs, plain Dirichlet smoothing seems to be the most appropriate choice since:

1. its probabilistic definition involves the simulation of a random QE.
2. it gives good results on short queries even using a non optimal μ parameter.
3. it favors short documents in focused documents search.
4. we are not considering verbose full text queries but queries combining concise MWTs.

Query Representation and ranking functions

Our purpose is to test the efficiency of MWTs in standard and focused retrieval compared to a classic bag-of-word model and statistically-derived phrases. For that, we shall consider phrases (instead of single terms) and a simple way of combining them. Given a phrase $s = (s_0, ..., s_n)$ and an integer k, we formally define the probability of finding the sequence s in the corpus with at most k insertions of terms in the following way. For any document D and integer k, we denote by $[s]_{D,k}$ the subset of $D_i \in D$ such that:

1. $D_i = s_1$
2. there exists n integers $i < x_1, ..., x_n \leq i + n + k$ such that for each $1 \leq j \leq n$ we have $s_j = D_{x_j}$.

We can now easily extend the definition of probabilities P and P_D to phrases s by setting $P(s) = P([s]_{.,k})$ and $P_D(s) = P_D([s]_{D,k})$. Therefore, we have:

$$P(s) = \frac{\sum_{q \in [s]_{D,k}, D \in \mathcal{D}} f_{q,D}}{|\Omega|} \tag{5.7}$$

$$P_D(s) = \frac{\sum_{q \in [s]_{D,k}} f_{q,D} + \mu \times P(s)}{|D| + \mu} \tag{5.8}$$

This is the easiest way to extend the usual multinomial model to phrases. Now, to consider queries that are set of phrases, we simply combine them using a weighted geometric mean as in Metzler and Croft (2003) for some sequence $w = (w_1, ..., w_n)$ of positive reals. Unless stated otherwise, we shall suppose that $w = (1, ..., 1)$, i.e. the normal geometric mean. Therefore, given a sequence of weighted phrases $Q = \{(s_1, w_1), ..., (s_n, w_n)\}$ as query, we shall rank documents acording to the following scoring function $\Delta_Q(D)$ defined by:

$$\Delta_Q(D) = \prod_{i=1}^{n} (P_D(s_i))^{\frac{w_i}{\sum_{j=1}^{n} w_j}} \tag{5.9}$$

$$\overset{\text{rank}}{=} \sum_{i=1}^{n} \left(\frac{w_i}{\sum_{j=1}^{n} w_j} \times \log(P_D(s_i)) \right) \tag{5.10}$$

This plain document ranking can easily be computed using any passage information retrevial engine. We chose for this purpose the Indri engine Strohman et al. (2005) since it combines a language model (LM) Ponte and Croft (1998) with an extention of

the INQuery language Callan et al. (1992) with a bayesian network approach which can handle very complex queries Metzler and Croft (2003). However, in our experiments, we use only a very small subset of the weighting and ranking functionalities available in Indri.

5.3.3 Query Expansion

We propose a simple QE process starting with an approximative short query $Q_{T,\mathcal{S}}$ of the form $(T, \mathcal{S})$ where $T = (t_1, ..., t_k)$ is an approximative document title consisting of a sequence of k words, followed by a possibly empty family of sets of phrases: $\mathcal{S} = \{S_1, ..., S_{|S|}\}$ where for each $1 \leq i \leq |S|$, S_i is of the form $\{S_{i,1}, ..., S_{i,l_i}\}$ for some $l_i \geq 0$. If $l_i = 0$ then S_i is considered to be the empty set. In our case, each $S_{i,j}$ will be a MWT.

Baseline document ranking fuction

By default, we shall rank documents acording to:

$$\Delta_{T,\mathcal{S}} \quad = \quad \Delta_T \times \prod_{i=1}^{|\mathcal{S}|} \prod_{j=1}^{|l_i|} \Delta_{S_{i,j}} \tag{5.11}$$

which is equivalent as ranking documents according to $\Delta_Q = \Delta_{T \cup \bigcup S}$ with the following weighting vector:

$$\left(\overbrace{\frac{1}{k}, ..., \frac{1}{k}}^{k}, \overbrace{\frac{1}{l_1}, ..., \frac{1}{l_1}}^{l_1}, ..., \overbrace{\frac{1}{l_\mathcal{S}}, ..., \frac{1}{l_\mathcal{S}}}^{l_\mathcal{S}} \right)$$

Therefore, the larger $\mathcal{S}$ is, the less the title part T is taken into account. Indeed, $\mathcal{S}$ consists of coherent subsets of MWTs defined by the user. If the user can expand the query by finding coherent clusters of terms, then we are no more in the situtation of a vague information need and documents should be first ranked according to precise MWTs. For our baseline, we shall generally consider $\mathcal{S}$ to be empty or made of phrases automatically generated from T.

Interactive Multiword Term Selection

The IQE process works in the following manner. We consider the top twenty ranked documents of Δ_Q ranking. The user selects a family $\mathcal{S}'$ of several subsets $S'_1, ..., S'_s$

of MWTs appearing in these documents. This leads to acquiring sets of synonyms, abbreviations, hypernyms, hyponyms and associated terms with which to expand the original query terms. We also let the user check that these terms do not introduce noise by adding them individually to the initial query and observing the top ranked documents. The selected multiword terms S'_i are added to the initial set $\mathcal{S}$ to form a new query $Q' = Q_{T,\mathcal{S}\cup\mathcal{S}'}$) leading to a new ranking $\Delta_{Q'}$ computed as in equation 5.11. We emphasize that $\mathcal{S}'$ is more than a flat list of MWTs. In our experiments we also evaluate if the structure of $\mathcal{S}'$ (i.e., grouping the MWTs into subsets) is relevant or not.

Automatic Query expansion

We also experimented with the automatic query expansion (AQE). In our model, it consists in the following. Let $D_1, ..., D_K$ be the top ranked documents by the initial query Q. Let $C = \cup_{i=1}^{K} D_i$ be the concatenation of these K top ranked documents. Terms c occuring in D can be ranked according to $P_C(c)$ as defined by equation (5.6). We consider the set E of the N terms $\{c_1, ..., c_N\}$ having the highest probability $P_C(c_i)$. We then consider the new ranking function Δ'_Q defined by:

$$\Delta'_Q \quad = \quad \Delta_Q^{\lambda} \times \Delta_E^{1-\lambda} \tag{5.12}$$

where $\lambda \in [0, 1]$.

Unless stated otherwise we shall take $K = 4$, $N = 50$ and $\lambda = 0.1$. We now explore in which context IQE based on MWTs is efficient. Our baseline is automatic document retrieval based on equation 5.9 in §5.3.2. We first show in §5.4 on the TREC Enterprise collections that this in fact is a very strong baseline. The results obtained on TRECEnt collections contrast somewhat with the very good results obtained by our IQE approach for the focused retrieval tasks on the Wikipedia corpus (see §5.5).

5.4 Enterprise search

As mentioned in section 5.2, it has been shown that language models based on Dirichlet smoothing as described in §5.3.3 are effective for retrieval from noisy large web collections especially with short queries Metzler et al. (2005). This probabilistic model relies on one single parameter: μ that can be viewed as the size of a pool in a survey. Therefore, whenever the size of the document collection is over $100,000$ documents, it can be observed that precision/recall functions do not significantly differ for $1000 < \mu < 2500$

as it has been pointed out in Zhai and Lafferty (2004). We show that this model also constitutes a strong baseline that is difficult to improve on smaller collections and on queries that are common in real corporate search environments.

5.4.1 Document retrieval at TREC-Enterprise track

The goal of the TREC enterprise track (TrecEnt) was "*to conduct experiments with enterprise data that reflect the experiences of users in real organizations*" Bailey et al. (2007). This track ran from 2004 to 2008. We participated in the 2008 edition but "trained" our search strategies beforehand on the 2007 data. Hence, we will indicate performances obtained on data from both years.

Document collection and Tasks

In 2007, the TrecEnt track chose the CSIRO Enterprise Research Collection (CERC) which is a crawl of all the `*.csiro.au` public websites performed in March 2007[2]. The collection consists of 370,715 documents totaling 4.2 gigabytes. The corpus contains approximately 7.9 million hyperlinks of which 95% pages have one or more outgoing links with anchor texts. However, the CSIRO collection differs from standard Web collections in that most links originate from the non-content part of the CSIRO pages. The search topics used in the TrecEnt tasks were furnished by employees of CSIRO in charge of science communication. These topics correspond to real world information needs received by the CSIRO staff from the public. Thus participating IR systems were judged on real life information needs and not on artificially contrived queries. Each topic consisted of two fields:

- a **query** field containing short query entities that the CSIRO staff would use to find information.
- a **narrative** field which is a substantive part of the e-mail.

The submitted runs were evaluated by the community based on the final answer furnished by CSIRO staff to the original requester. Figure 5.1 gives an example of a topic from TrecEnt 2008.

[2]the Australian 'Commonwealth Scientific and Industrial Research Organization'

<top>
<num>CE-051</num>
<query>weatherwall</query>
<narr>Have been trying to access the CSIRO weatherwall site to check on weather in Melbourne over the last 24 hours. It seems to be off line at present. Any idea why? When might it be back on line? </narr>
</top>

Figure 5.1: Example of a topic in the TRECEnt 2008 track.

In a real life situation, this information request will involve a humain to not only point to the appropriate web page (weatherwall) but also to respond to actual information need. Such information requests were not in the minority in the topics delivered to TrecEnt participants. The best automatic IR systems can do is to point to pages containing some terms on the object of discourse (here weatherwall page) but they cannot respond to the actual information need here which is "why did the weatherwall service break down and when it be functioning again"?

Description of runs

First, the CSIRO corpus was indexed. We applied the Porter's stemmer implemented in Lemur toolkit[3] in order to acquire more word frequencies irrespective of inflection. However, we did not use any stop word list. We designed four basic search strategies, called "runs" in the TREC terminology. These four runs will be applied on the 2007 and 2008 TrecEnt collections as well on the INEX Ad-hoc tasks albeit with some variations. The first run is the baseline defined in §5.3.3 using only the query fields. The second is a boosting of this baseline by simply repeating queries in the $\mathcal{S}$ component as phrases. Clearly, instead of leaving $\mathcal{S}$ empty in equation 5.11, $\mathcal{S}$ is the singleton $\{\{q\}\}$ made of the query phrase q. The last two runs are based on the IQE process described in 5.3.3 and published in SanJuan et al. (2008). We give below the precise details of each run:

- **baseline bag-of-words (baseline-B)**: we set $T = \{q_1, ..., q_n\}$ where the q_i are the terms in topic query field q. $\mathcal{S}$ is left empty. This is the usual multinomial bag-of-word approach.
- **baseline phrases (baseline-P)**: we keep the same T but $\mathcal{S}$ is set to the singleton $\{\{(q_1, ..., q_n)\}\}$ whenever the query contains at least two words, i.e. in addition to the bag-of-words approach, we also consider the query q as a phrase.

[3] `http://www.lemurproject.org/`

- **IQE MWT-groupings (IQE-C)**: this run corresponds to the IQE approach described in §5.3.3 except that the user creates sub-groups of MWTs, hence providing a hierarchy of sorts among MWTs. We set $\mathcal{S}$ to $\mathcal{S}(t)$ for each topic. The T component is unchanged.

- **IQE MWTs flat list (IQE-L)**: we consider as $\mathcal{S}$ a flat version of each $\mathcal{S}_t$:

$$\mathcal{S} \quad = \quad \left\{\bigcup \mathcal{S}(t)\right\} = \{\{t : t \in S_i, S_i \in \mathcal{S}(t)\}\} \tag{5.13}$$

  where all the selected MWTs are considered at the same level, the internal structure of $\mathcal{S}(t)$ is ignored.

The $\mathbf{IQE-L}$ run evaluates the impact of $MWTs$ on document ranking while the **IQE-C** run, also based on MWTs, evaluates the impact on the retrieval effectiveness of forming subsets of $MWTs$ by the user. We illustrate these two representations of MWTs on the same topic as in figure 5.1. For the **IQE-C run**, the user formed these subsets of MWT queries:

1. {weatherwall}
2. {(weatherwall site), weather, Melbourne}
3. {(CSIRO weatherwall site), weatherwall, (weather in Melbourne)}

In this representation, the particular angle by which the MWT is sought is reflected by a facet term placed to the right of it, e.g. *((weatherwall site), weather, Melbourne))*. In the **IQE-L run**, the expanded query is represented by this flat list of MWTs: *((weatherwall site), (CSIRO weatherwall site), (weather in Melbourne), weatherwall, weather, Melbourne*). This is a simplified version of the same MWTs used in the IQE-C run in which the facet terms have been removed. All terms are weighted equally here.

5.4.2 Results based on usual Average Precision

The official measure for the TrecEnt 2007 edition was Average Precision (AP). This was changed to *inferred* Average Precision (*inf*AP) for TRECEnt 2008. However, we can compute AP on both tracks.

Document search on the TrecEnt 2007 collection

50 topics were provided and all were judged. On the resulting document qrels, our baseline reaches a mean average precision (MAP) of 0.441 which outperforms all reported runs in Bailey et al. (2007), the highest MAP being 0.422. However, based on the query by query average precision (AP) score, there is no statistical evidence (t-test with a 95% confidence interval) that our baseline has a true mean not equal to 0.422. Since TrecEnt queries were short phrases most of which had the appearance of MWTs like *"solve magazine, selenium soil"*, the question was to ascertain if our baseline can be boosted by considering phases as suggested by Mishne and de Rijke (2006). It seems the answer is yes, but only slightly since the *phrases* run (see 5.4.1) reaches the MAP score of 0.448. However, this improvement is not statistically significant. The corresponding Recall/Precision curves on 2007 queries are shown in figure 5.2 together with curves on 2008 queries that we discuss in the next subsection.

Document search on the TrecEnt 2008 collection

77 topics were made available to participants of which 67 were judged. Four had no judged relevant documents and were dropped. For our participation to this track San-Juan et al. (2008), the same IQE process was implemented in which a user selected for each topic t, subsets $\mathcal{S}(t)$ of MWTs following the methodology described in §5.3.3.

We first computed the AP measures used in TrecEnt 2007 in other to compare our baseline to its performance on this data. Confirming its good perfomance in 2007, our *baseline-B* run implementing the bag-of-word approach outperformed all our other approaches. The good performance of our *baseline-B* here confirms that it is indeed a strong one since it reaches similar precision scores at 10% of recall and even higher at 20% of recall as illustrated in figure 5.2. The 2008 curves then drop because TrecEnt 2008 qrels are based on a more complex pooling process that handicaps low ranked documents in participant runs. Indeed, AP was not the official measure used in TrecEnt 2008 but inferred AP as described in Bailey et al. (2008). Before detailing in §5.4.3 reasons for this change of measure, let us observe on the same figure the two curves induced by IQE runs on 2008 queries. It appears that only IQE-C run succeeds in slightly improving the baseline line runs at 5% of recall, but then, like the other IQE-L run, it drops under the baseline.

In fact, it appears that our two baseline runs ranked first the "easiest to find" relevant documents among these qrels. These are documents found by most participants.

Therefore, to have an insight look in system performance on such large and noisy enterprise corpus, it is necessary to take into consideration the probability of picking a document at a given rank from participant runs. This explains why in TrecEnt 2008, the official metric used for evaluating participating systems for this task was not based on simple average interpolated precision but an inferred measure based on a stratified sample.

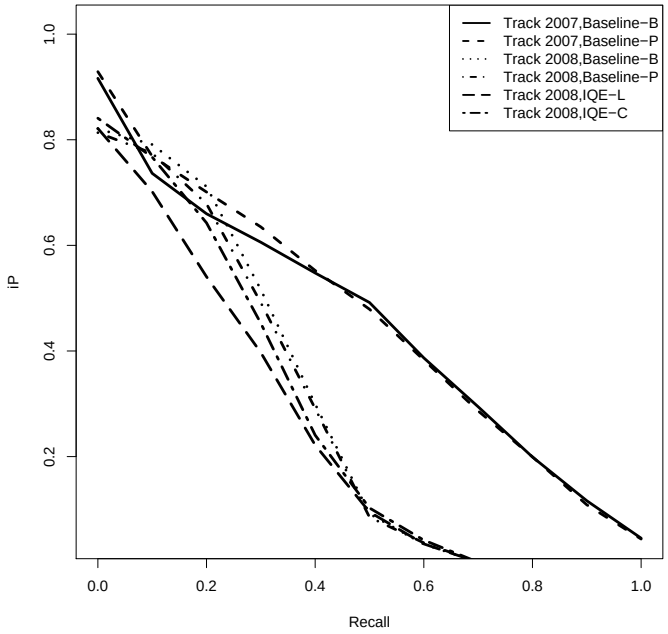

Figure 5.2: Absolute Precision/Recall curves computed on TrecEnt 2007 qrels and 2008 qrels without considering available sampling information.

5.4.3 Results based on Inferred Average Precision

Usually in the qrels, for a document to be judged relevant it has to be:

- in the pool of documents selected among runs submitted by participants,
- marked as relevant by at least two assessors

The inferred AP (infAP) measure used in TRECEnt 2008 is similar to the original infAP used in the TREC Terabyte track, except that it has been modified to work on stratified samples. Both versions of infAP take into account the fact that the measurement is based on a pool of relevant documents and not on an exhaustive list of all

relevant documents. Indeed, AP relies on the knowledge of the complete set of relevant documents which on a large corpus is not generally known. According to NIST organizers of the TrecEnt 2008, *"two runs were pooled out from each group to depth 100. The documents were selected for judging by taking a stratified sample of that pool based on document ranks: documents retrieved at ranks 1-3 were sampled at 100% depth, documents of ranks 4-25 at depth 20%, and document between 25-75 rank were sampled at 10% depth. The rank of a document for sampling purposes is the highest rank over all pooled runs. "* The evaluation script and relevance judgments are available from the TREC website[4]. The script also allows us to estimate the usual Normalized Discounted Cumulated Gain (NDCG) that gives more importance to elements at higher ranks. Figure 5.3 shows the inferred AP and NDCG of our baseline and IQE runs.

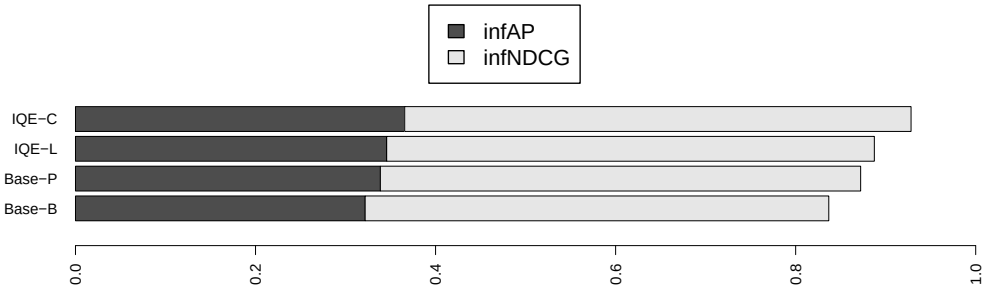

Figure 5.3: Inferred Average Precision and Normalized Discounted Cumulated Gain on TrecEnt 2008 qrels using available sampling information.

On the resulting 2008 stratified qrels, our *baseline-B* run attains an infAP score of 0.3218 thus placing itself among the six best runs submitted to TrecEnt 2008. Our best submitted run at TrecEnt 2008 was ranked 7th SanJuan et al. (2008) but it was not based on the same language model explored here. In contrast with previous results on absolute AP, the infAP goes up to 0.3387 when considering phrases in *baseline-P* run, 0.345 when considering *IQE-L* run based on the flat list of additional terms and 0.3657 for *IQE-C* run using the grouped set $\mathcal{S}(t)$ of MWTs. Therefore, using the infAP measure, our IQE-MWTs runs outperform the baseline bag-of-word and phrase runs.

However, only the difference between the first *baseline-B* and other runs is statistically significant (t-test at 95% of confidence). Other differences are not significant. Since the *baseline-P* run is in fact the *baseline-B* boosted by adding the whole topic query as a phrase to the initial bag of words query, these results show that Mishne and de Rijke (2006)'s observations that document retrieval performance can be boosted on large web collections by considering phrases, are also true on smaller enterprise web corpus. Based on the results from TRECEnt 2007 and 2008 data, we cannot infer that

[4] `http://http://trec.nist.gov/data/t17_enterprise.html/`

IQE based on MWTs brings significant improvement in document search, even if, like in Vechtomova (2005), we do observe some improvement. We need to further confirm the rather conflicting results obtained on the CSIRO collection which may be due to the change from average precision (AP) to inferred average precision (infAP) between TRECEnt 2007 and 2008 editions. We also hypothesize that differences in collection quality played a significant role (CSIRO web domain vs Wikipedia).

5.5 Focused retrieval

The focused retrieval experiment was carried out in the framework of INEX 2008 Ad-hoc track which is the main forum for researchers working on the extraction of information from structured documents, mostly XML Lalmas and Tombros (2007). Given the prevalence of XML in electronic information systems, being able to locate the specific part of a text that is relevant to a user's query is in line with the current research on question-answering. XML offers the opportunity to exploit the internal structure of documents in order to allow for more precise access, thus providing more specific answers to user requests.

Of particular interest to us is the fact that passage or element retrieval from structured documents has required the development and the experimentation of new measures to evaluate focused retrieval Gövert et al. (2006); Kamps et al. (2007). These measures induce a new approach to document relevance. For instance, when searching an encyclopedia, these measures will favour the ranking of articles that answer the query exhaustively, against some articles that contain some passages related to the query. Since each article in an encyclopedia is usually about a specific topic and does not digress in its contents, this tends to support the relevance of full article retrieval in INEX's focused retrieval task. Hence, the INEX ad-hoc task using Wikipedia as test collection also offers a convenient framework to evaluate standard document retrieval systems that favour short but entirely relevant documents. Observations reported in Mishne and de Rijke (2006) on phrase queries suggest that IQE based on MWTs should render good results on such document collections. This is precisely the object of our experiment here. Indeed the results obtained on the Wikipedia corpus tend to confirm Mishne and de Rijke (2006)'s assumption.

5.5.1 INEX 2008 Ad-hoc track

Corpus and topics

The official INEX 2008 corpus was the 2006 version of the English Wikipedia comprising 659,388 articles without images L. Denoyer (2006). On average, an article contains 161 XML nodes, where the average depth of a node in the XML tree of the document is 6.72. From this corpus, participants were asked to submit query topics corresponding to real life information needs. A total of 135 such topics were built, numbered from 544-678. 70 out of them were judged by the community and thus used in the official evaluation. A topic consists of four fields: content only field (<CO> or <Title>) with a multi-word term expression of the topic; a content only + structure version of the topic (<CAS>) which is the title with indication of XML structure where the relevant elements may be found; a <description> field which is a slightly longer version of the title field; and a <narrative> field comprising a summary with more details about the expected answers. Typically, the narrative would indicate things to eliminate from relevant documents and draw boundaries that can be geographic, spatial, genre or historical in nature. Some title fields contained Boolean operators that required systems to explicitly exclude (-) or include (+) certain terms in the relevant answer elements.

Ad-Hoc Retrieval Tasks

The 2008 Ad-Hoc track had 3 tasks: Focused retrieval, Relevant-in-Context (RiC), Best-in-Context (BiC).

1. The focused task requires systems to return a ranked list of relevant non-overlapping elements or passages. This is called the "fetching phase".

2. The Relevant-in-Context (RiC) task builds on the results of the focused task. This task is based on the assumption that a relevant article will likely contain relevant information that could be spread across different elements. This is called the "browsing phase". Systems are therefore asked to select, within relevant articles, several non-overlapping elements or passages that are specifically relevant to the topic.

3. The Best-in-Context (BiC) task is aimed at identifying the best entry point (BEP) to start reading a relevant article. This task is based on the assumption that "even

an article completely devoted to the topic of request will only have one best starting point from which to read (even if that is the beginning of the article)" Kamps et al. (2008).

Extended qrels and evaluation measures

The evaluation procedure establishes an extended qrel file similar to those used in TREC against which all participating systems are evaluated. Like in TREC Terabyte and Adhoc tracks, the procedure consists in selecting for each query a pool of documents from participant runs. Topics and documents are then randomly distributed to assessors from the INEX community. Using an ergonomic java on-line interface, each assessor has to mark-up for each document, the relevant passages with regard to a topic. It is important to emphasize that query terms are highlighted in the display of documents. Moreover, in 2008, the interface offered the facility of selecting the whole document using a simple radio button. The assessor had also to point out the BEP. These result in a qrel file that gives for each evaluated pair of topic and document, the total length of relevant passages, the document length, the offset of the BEP and the list of relevant passages. Lengths are computed as number of characters in the text version of the corpus (without XML tags). The 2008 qrel file required the evaluation of $36,605$ articles. Among them, only $4,773$ were judged to contain at least one relevant passage for at least one topic. However, it appears that 40% of these $4,773$ documents have at least 95% of their content marked as relevant by assessors. These highly relevant documents only cover 0.02% of the total length of evaluated documents but almost 25% of the total length of relevant passages. These facts are important to estimate the upper AP bound for systems retrieving full document instead of passages or XML elements. Indeed, following these qrels, for focused task, precision is computed as the total length of relevant sub-passages in the ranked documents over the total length of retrieved passages Kamps et al. (2008). Similarly, recall is computed taking the same length of relevant sub-passages, divided by the total length of relevant documents. Following this definition, we found out that based on 2008 INEX qrels, a system that retrieved first almost all completely relevant documents can reach an expected total precision of 25% and an AP of 22%. Therefore, INEX 2008 focused task was the ideal framework to evaluate such systems.

The RiC and BiC are also evaluated based on these qrels but using graded document scores whereas in the focused task, scores are based on the sole relevant passages no matter their co-occurrence in documents. Given a document score function S into $[0, 1]$, both RiC and BiC evaluations are based on generalized precision gP at some rank r which is the average score S over the r scores documents. Given a document d, the

score $S(d)$ is in the case of:

- RiC, the F-score of the retrieved passages from d by the system among all relevant passages in d.
- BiC, a normalized distance in number of characters between the BEP found by the system and the real one.

The consequence is that these measures favour even more full document retrieval strategies against passage retrieval since for 40% of relevant documents, full document retrieval strategies will obtain the maximal score whenever they retrieve relevant documents. We refer to Kamps et al. (2008) for further discussion of these measures.

5.5.2 Results

We first present our search strategies, then analyze results by tasks in the INEX Ad-hoc track.

Runs

We consider the same four basic strategies as in the TREC Enterprise search track (see 5.4.1): *baseline bag-of-words (baseline-B), baseline phrases (baseline-P), IQE MWTs subsets (IQE-C)* and *IQE MWT flat list (IQE-L)*. Like in the TrecEnt experiment, the two first runs are automatic, the last two rely on the sets of MWTs manually gathered when browsing the top ranked 20 documents based on an initial query. Table 5.1 gives an example of such expansion.

IQE-LC with subsets of MWTs	resulting flat list for IQE-C
{(dna testing) disease}	(dna testing)
{(dna testing ancestry)}	(dna testing ancestry)
{(genetic disease), (dna testing) ancestry}	(genetic disease)
{(hereditary disease) (dna testing) ancestry}	(hereditary disease)

Table 5.1: Selected multiword terms for the INEX 2008 topic “dna testing forensic maternity paternity".

Compared to the TrecEnt runs, there are two differences in the way that we apply these runs here:

1. we do not use any stemmer, nor lemmatization and we index all the text (no stop word list).

2. we systematically apply AQE to all runs.

Indeed, we observed in earlier experiments Ibekwe and SanJuan (2009) that by combining the above two complementary features, results for the Wikipedia corpus are significantly boosted whereas on the CSIRO web collection, these two features had the opposite effect. This can be easily explained by the fact that Wikipedia articles are well written, with very few spelling errors, thus any stemming will induce a loss of information whereas on the CSIRO web pages, stemming tended to reduce the noise. AQE on the non lemmatized Wikipedia corpus was able to automatically capture synonyms and some grammatical variants of the query term. On the CSIRO corpus used in TrecEnt, AQE just added more noise.

Focused task

The INEX 2008 official measure for focused task was average interpolated Precision at 1% of recall (iP[0.01]). Figure 5.4 shows the Recall/Precision curves of our baseline and IQE runs. The best score for all runs in the official evaluation was 0.6896. Our *baseline-B* score (automatic run with AQE) obtains a significantly much lower score at 0.5737. The *baseline-P* run did not benefit from the same boosting effect as in TRECEnt experiment, hence its much lower score of 0.5732. The *IQE-L* run obtained a much higher score of 0.7016, even higher than the best participating system. This score is further improved to 0.7137 when we consider the *IQE-C* run in which MWTs had been grouped to reflect more complex query representations (see table 5.1 for an example).

The differences between IQE-based runs are not statistically significant, whereas the difference between *baseline* runs and the IQE runs is this time clearly significant. Indeed, using the Welch Two Sample paired t-test, we find a p-value of 0.02302. Moreover, other participants' best runs submitted at INEX 2008 were optimal for very low recall values but then drop down fast for higher recall values. One might put forward the argument that the good score of our IQE runs may be due to the fact that the user found one or two completely relevant documents with some specific MWTs which were then re-introduced in the expanded query. The Precision/Recall curves in Figure 5.4 show that this was not the case. In fact, mean average iP for the *baseline* runs is only 0.28 while that of both both IQE runs reach 0.34. The difference is again statistically significant at 95% of confidence with an estimated p-value of 0.03966. Therefore, this

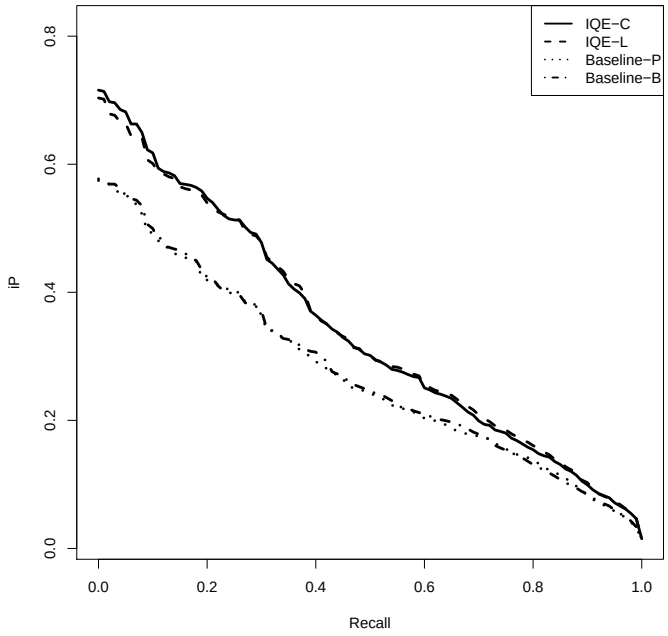

Figure 5.4: Focused interpolated precision curves on INEX 2008 topics.

experiment clearly demonstrates that representing queries with MWTs corresponding to real concepts instead of n-grams or bag-of-words, can dramatically improve IR when dealing with a high quality collection such as the Wikipedia. We now present results for the other two tasks of the Ad-hoc track.

Relevant-in-Context and Best-in-Context tasks

The official measure for these tasks was MAgP (Mean Average generalized Precision). By considering that we only retrieve articles that are completely relevant, and that the best entry point is the first character of the document, the same four runs can be evaluated with regard to the RiC and BiC measures.

Our runs maintained the same order as it can be observed in figure 5.5. Among all submitted runs to INEX 2008, the best score was 0.228 for RiC and 0.224 for BiC. Our *baseline* already reaches a score of 0.197 for RiC and 0.20 for BiC. This places our baseline among the six best runs and our group among the three best teams. The baseline is slightly improved by considering *phrases*: 0.2 for RiC, 0.206 for BiC. The scores of IQE outperform the best scores in the official evaluation. Indeed, the *IQE-L* run reaches a score of 0.236 for RiC and 0.248 for BiC. Surprisingly, *IQE-C* run does not

improve these score since it obtains a score of 0.235 for RiC and 0.246 for BiC. However, none of these differences are statistically significant at 95% of confidence, the Welch Two Sample t-test p-value between the *baseline* and the *IQE-L* runs being 0.08739 for RiC and 0.05981 for BiC. Classical MAP was also computed at INEX 2008 by considering as relevant any document involving at least one relevant passage, whatever its length. There, we also find that IQE runs also outperform all other runs, but the difference with the baseline is even less significant.

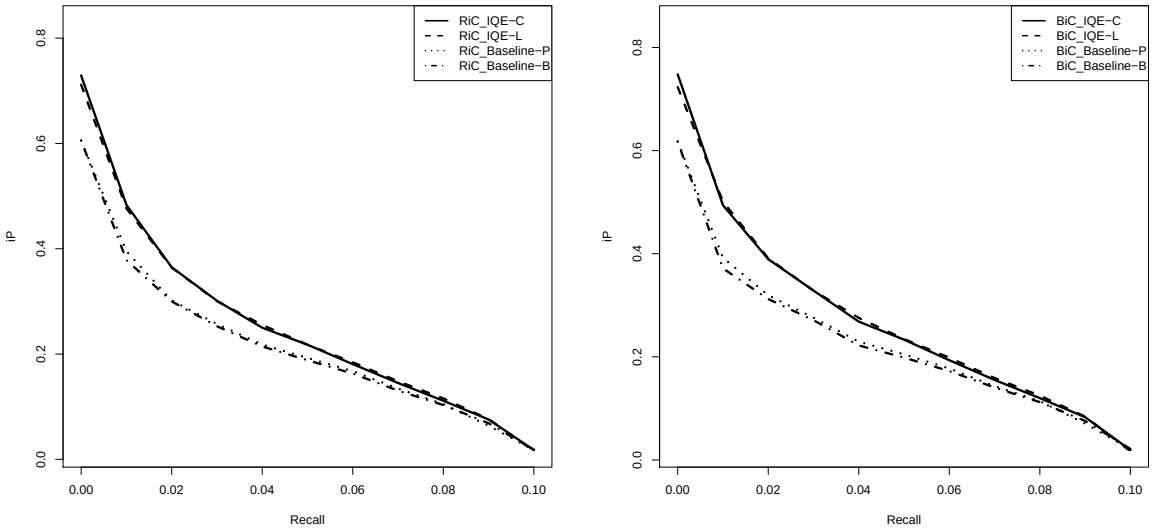

Figure 5.5: Interpolated generalized precision curves on INEX 2008 topics for Relevant in Context (left) and Best in Context (right).

We further tested our search strategies on an element retrieval algorithm in order to retrieve XML elements or passages instead of full documents.

Element retrieval

It follows from previous results that the focused measure clearly shows the improvement that can be expected using an IQE procedure based on MWTs. However, the focused measure was not initially meant to evaluate full document retrieval (fdR), but XML element retrieval (xmlR). Even though our IQE fdR runs outperform all xmlR runs submitted to INEX 2008, we have investigated what happens if we use the same IQE process for xmlR. We applied the following strategy. We extract per topic all relevant paragraphs (tag p), sections and articles. Then, for each relevant article, we check if there is a disjoint list of relevant paragraphs. If there is, we return these paragraphs

	Baseline-B	Baseline-P	IQE-L	IQE-C
TrecEnt MiAP	0.322	0.339*	0.345	**0.366**
INEX focused iP[0.01]	0.574	0.573	0.702*	**0.714**
INEX focused MAiP	0.280	0.280	**0.340***	**0.340**
INEX RiC MAgP	0.197	0.200	**0.236**	0.235
INEX BiC MAgP	0.200	0.206	**0.248**	0.246

Table 5.2: Summary of results between the four runs over the two corpus TrecEnt and INEX 2008 . Figures marked with * are statistically significantly greater than lower figures on the same row. Best scores are in bold form.

instead of the whole document. If there is no paragraph, we try the same procedure with sections. If no paragraphs and no sections are found, we return the whole article. The results are that only scores at iP[0.00] are improved, then they immediately drop down. The IQE scores are still outperforming the baseline but not in a significant way.

5.6 Discussion

Table 5.2 recalls the main results that we obtained on TrecEnt and INEX collections.

This table shows that at least on focused retrieval task, IQE runs based on MWTs selected from top n-ranked documents significantly improves automatic state of art IR. It also outperformed state-of-the-art systems for passage retrieval because, as expected, even if in our experiments we only retrieve entire documents, our methodology focused on documents that are entirely relevant to the query and due to the encyclopedic organization and exhaustiveness of the Wikipedia, often the most relevant passage to a query is the whole document. However, there is no clear evidence that IQE based on a grouping of MWTs into subsets (IQE-C run) outperforms simple IQE-L based on a flat list of MWTs. This is interesting in the perspective of automating the MWT selection phase. Like we mentioned in section 5.3.1, on the linguistic level, terms are a subgroup of phrases (mostly noun phrases) that correspond to domain concepts. Hence, it is not always possible to distinguish them from ordinary noun phrases based solely on their morpho-syntactic composition. What most term extraction tools extract are candidate terms and it is often left to humans to distinguish real domain terms from general noun phrases. At this initial and exploratory phase of methodology, we wanted to first ascertain that MWTs are indeed more effective than n-grams, bag-of-words or general phrases before automating their extraction in order to assist the IQE process.

Now that this has been ascertained to a degree to which it is an interesting option, we have tools that can extract the MWTs. In earlier research, we have built a text clustering system (TermWatch) that first extracts MWTs from terms before clustering and mapping them SanJuan and Ibekwe-Sanjuan (2006). To determine whether the MWTs that were manually selected in the IR experiments reported here can be acquired by our system, we subjected the ones selected in the INEX Ad-hoc runs to our text mining system.

The complete set of 135 title topics built in INEX 2008 are themselves MWTs. 719 supplementary MWTs were manually selected from top n-ranked articles from the Wikipedia following an initial query. The number of selected MWTs per topic range between 1 and 20, the average number being 5.30. Each MWT has a length between 1 and 9 words (nouns, adjectives or prepositions), the average length being 2.70. We studied how these terms are related to the topic titles and descriptions by uploading them into the TermWatch system. TermWatch is a text mining platform integrating NLP techniques and clustering algorithm. The system can cluster MWTs based on different levels of linguistic variations. Here we only used syntactic variations to establish relations between the selected MWTs. In this case, two terms will be related if one can be derived from the other by adding or inserting a sequence of words at a single position, or by substituting one word by another. It appeared that among the 719 manually selected MWTs, 206 are related to the title field by a chain of such variations. 110 other terms are related to MWTs in the description field of the topic. Therefore, 44% of the manually selected MWTs were variants of MWTs in the topic description, found in the top n articles returned from an initial query. Since the set of user-selected MWTs in our IQE process are in most part related by syntactic variations to the MWTs in the topic description, they can be acquired automatically by TermWatch or other NLP tools. However, not all syntactic variants are relevant and user feedback may *in fine* still be needed to weed out the indesirable ones as done in (Vechtomova (2005)). Based on these findings, we devised a supplementary run by expanding the baseline run with these 316 related MWTs. This run obtained an IP[0.01] of 0.627 and a MAiP of 0.296. These scores are significantly better (with 90% of confidence) than the baseline, but also significantly different from the best ones, including complete IQE scores.

This last experiment shows that such subset of syntacticallly related MWTs is not sufficient and it is necessary to consider syntactically unrelated MWTs.

Another issue which may represent a bias in our experiments is that the manually selected MWTs used for IQE runs was performed by a single user. This is not unsual for exploratory methodological studies. In the extensive study by Perez-Carballo and Strzalkowski (2000), a single user, in fact one of the authors of the paper carried out the

IQE experiments. The reasons being that before expending much effort on designing interactive search interfaces for a panel of users, one must first ascertain the potential effectiveness of the IQE process against wholly automated procedures which are already "around". Hence, it is possible that our single user here, being also one of the authors of the paper is more sensitive to common pitfalls of search techniques and was more able to select better quality terms than average *lambda* users in real life situations. This is a possible bias whose impact will be investigated in further studies. However, we can safely put forward the argument that by profiling the users' level of knowledge with search processes, i.e., using undergraduates or graduate students who are familiar with searching different data repositories, the gap between the quality of MWTs selected in this study and the future ones will not be too significant.

5.7 List of MWTs used for the 20 first TREC Enterprise 2008 topics

(51) weatherwall: CSIRO weatherwall site; weather in Melbourne; weatherwall site;
(52) solve magazine: solve magazine; solve website; solve magazine;
(53) selenium soil: selenium fertilisers; selenium deficiency; selenium response; selenium supplementation; selenium soil; soil additive;
(54) the heat is on: the heat is on; energy report;
(55) case moth identification: case moth;
(56) 12345 : 1235 Food and Nutrition Plan;
(57) fast instruments: FAST instruments; axon instruments; screening pharmaceutical drugs;
(58) wood borer treatment: wood borer; pest management; pest control; wood borer treatment; lyctid borers; borer fluid; infested timber;
(59) vinelogic cd: vinelogic; cd rom; vinelogic education;
(60) algae hydrogen powered cars: automotive transport; hybrid electric vehicles; ecommodore; hybrid family cars; low emissions vehicle project; hydrogen powered fuel cells; energy saving vehicle; hybrid vehicle; clean car;
(61) wheel motor: wheel motor; solar cars; aurora solar car; solar racing; sun powered cars;
(62) climate change hops: hop cultivar; beer industry; australian bred hops; hops industry; climate change; hop cultivar; save the ales;
(63) vinegar bugs: vinegar bugs; vinegar fly; vinegar flies;
(64) ant identification: ant identification; shiny blue ant; australian ants online website;
(65) recruitment: career prospects; employment conditions; people and skills; job vacancies;

(66) chilli paste preservatives: food preservative; food conservation; spice preservative;
(67) sydney ocean temperatures: Sydney coastal waters; new south wales coast; new south wales coast;
(68) sea level changes east coast: sea level change; sea level rise; satellite almitery; East Coast; gold Coast; sea level changes; East Coast;
(69) magnesium supplement: magnesium supplement; oral supplement of magnesium; mg supplementation;
(70) information technology jobs: information technology positions; positions vacant; career online; systems administrator; software engineer; digital systems engineer; software developer; applications developer; job opportunity; employment opportunity; vacancies; systems administrator; software engineer programmer; digital systems engineer; software engineer; applications developer;
(71) termite wings: white ants; life cycle; termite wings; white ants; life cycle termite;

5.8 List of MWTs used for the 20 first INEX 2008 ad-hoc topics

(544) meaning of life: center of life; direction of life; existence; life wheel; meaning of life; nature of life; philosopher of life; philosophy of existence; philosophy of life; purpose life; reflection of life; significance of life; source of life;
(545) dance style: ballroom dancing; body contact dance; dance improvisation; dance style; dance technique; dance technology; dancing style; folk dance; folk dances; lead and follow dance; list of dances; list of dance style categories; list of novelty dances;
(546) 19th century imperialism: 19th century; 19th century imperialism; colonial empire; new imperialism;
(547) greek revolution 1821: greco turkish war; greek war of independence;
(548) health insurance policy national: canada health act; health care financing; health insurance; health insurance fund; health insurance reform; long term care insurance; medical insurance; national health insurance system; primary health care; public health system; publicly funded health care; publicly funded medicare; publicly funded medicine; social health insurance; State health insurance; universal health coverage;
(549) anti aging treatment: aging association; anti aging medicine; anti aging treatment; aubrey de grey; cellular aging; engineered negligible senescence; extend lifespan; human aging; human lifespan; life extension; retardation of aging;
(550) dna testing forensic maternity paternity: dna testing; dna testing ancestry; genetic disease; hereditary disease;
(551) pollen allergy: allergic rhinitis; hay fever; nasal allergies; pollen allergy;

(552) keyboard instrument electronic: keyboard instrument; keyboard stringed instrument; string instrument; wind instrument;
(553) spanish classical guitar players: classical guitarist; spanish classical guitarist;
(554) barney and friends: barney and friends;
(555) amsterdam picture image: image of amsterdam; images of amsterdam;
(556) vegetarian person she woman: list of vegetarians; notable vegetarians;
(557) electromagnetic waves: electromagnetic radiation; electromagnetic waves; gamma rays; infrared radiation; microwaves; radio waves; terahertz radiation; ultraviolet radiation; X rays;
(558) ufo sight places: Cash Landrum incident; Height 611 UFO incident; Lonnie Zamora; Project Blue Book; Rendlesham Forest Incident; UFO incident; UFO reports; UFO sightings; unidentified flying objects;
(559) vodka producing countries: Absolut Vodka; belvedere vodka; grey goose vodka; iceberg vodka; ketel one vodka; koskenkorva vodka; list of vodkas; pearl vodka; shakers vodka; siwucha vodka; skyy vodka; smirnoff vodka; van gogh vodka; vladivar vodka; vodka 1; vodka oso negro; vodka production; wisla vodka; xellent swiss vodka; zyr russian vodka;
(560) european cities with skycrapers higher than 100 meters: 25 Canada Square London; 8 Canada Square London; Commerzbank Tower Frankfurt; european skyscrapers; Federation Tower; Messeturm skyscraper; Millennium Tower Vienna; One Canada Square; Palace of Culture and Science Warsaw; Skyscrapers in Europe; Skyscrapers in London; Skyscrapers in Poland; tallest building in the UK; tallest buildings in Europe; Torre Espacio; Tour Montparnasse paris; Triumph Palace Moscow; Warsaw Trade Tower;
(561) portuguese typical dishes: arroz doce; bolinhos de bacalhau; caldo verde; portuguese bacalhau; portuguese cuisine; Portuguese cuisine; sardinhas assadas;
(562) algerian war: Ahmed Ben Bella; algerian war; Algerian War of Independence; Algiers putsch; battle of algiers; evian Accords; National Liberation Front ; Paris massacre of 1961; Pied Noir; secret army organization;
(563) virginia woolf novels: Mrs Dalloway; Night and Day; The Voyage Out; The Waves; To the Lighthouse; virginia woolf; Virginia Woolf;
(564) criticism limitation null hypothesis significance test: alternative hypothesis; null hypothesis; null hypothesis testing; type I error; type II error;

5.9 Conclusions

We have presented in this chapter a methodology that relies on meaningful text units (multiword terms) to represent queries. These multiword terms are used alternatively with interactive query expansion and automatic query expansion, the two are also combined in order to determine the combination that best boosts retrieval effectiveness.

The experimentation has been carried out on two different document collections: a web collection consisting of the CSIRO domain and the Wikipedia corpus within TREC Enterprise track and INEX Ad-hoc track respectively. While the results obtained on the TrecEnt collection are not conclusive due perhaps to poor corpus quality and a change of evaluation measures in the TrecEnt campaigns, the results on the Wikipedia collection show that multiword term query representation and interactive query expansion are a promising combination for both standard document and focused retrieval.

Chapter 6

Microblog Contextualization: setting up a new game combining focused retrieval and automatic summarization

6.1 Sentence Ranking

All previous experiments involved few players. Here we set up an open INEX track around a challenging task. A baseline system implementing previous results from chapters 4 and 3 is proposed. Performance is based evaluated based on informativeness and readability.

The first automatic multi-document summary systems appeared in the 90's McKeown and Radev (1995). Most of them apply statistical techniques to text segments like terms, sentences etc. to rank them according their relevance Mani and Mayburi (1999). The abstract is then generated by extraction of the top ranked sentences among all documents.

We consider a variant of TextRank algorithm Mihalcea (2004). Indeed, TextRank algorithm can be restated in terms of matrix S and E. Given a matrix M, let us denote by $D(m)$ its diagonal. TextRank computes a sequence of matrices $R_0, ..., R_n, ...$ where:

$$R_0 = D(E) + S \times S^t - D(S \times S^t) \quad (6.1)$$

$$R_{n+1} = D(R_n^2) + R_n - D(R_n) \quad (6.2)$$

The algorithm is iterated until $|D(R_n^2) - D(R_n)| < \epsilon$ for some fixed ϵ. The idea of TextRank is to progressively weight the sentences according to the number of sentences in their neighborhood in graph $G_{S^t \times S}$ and on the strength of an edge between two sentences as the number of common words. The algorithm converges towards the same solution no matter the initial weights on vertices. We clearly have that for $n \leq 3$, $D(R_3) \leq D(E)$.

The analogy with magnetism systems Fernández et al. (2007) suggest to directly consider matrix E instead of R_n. We have check on small toy samples that the two methods produce the same ranking. The rankings differ on large samples whenever the algorithm needs to be iterated more than 30 times.

The rest of this section is devoted to show that E is sufficient to rank sentences in a automatic summary perspective. The experiment is carried out on DUC data and results are compared to NUS system Lin et al. (2007) that uses TextRank. Document Understanding Conferences (DUC) were organized from 2001 to 2007 by NIST[1]. The main task of DUC consisted in handling complex and real questions. The resulting answer cannot be simple entity (name, date or quantity as this is usually the case in Quesion-Answering (QA) track in TREC conference[2]). The problem can be stated this way. Given a topic and a set D of relevant documents, generate a short correct and coherent summary of 250 words, that will answer to questions in the topic. Topics are made of two parts: a title and a short description. The $|D = 25|$ documents were extracted from AQUAINT corpus made of news from *Associated Press*, *New York Times* (1998-2000) and *Xinhua News Agency* (1996-2000).

Summarizers based in sentence extraction, introduce the query as a supplementary sentence. The sentences are extracted from documents according to their distance to query. In the case of TextRank approaches, the query is introduced as a supplementary graph vertex. In TextRank approach, a sentence i is ranked according to its score $[R_n]_{i,i}$. In textual energy approach it is ranked according to $E_{i,q}$ where q corresponds to the query.

Systems then select the top ranked sentences such that the total number of words is less than 255. Thus, the number of selected sentences changes according to their length. Sentences are then displayed in an order that respects the one of their appearance in documents. Moreover, to avoid redundancy, which is an important question in multi-document summarization, sentences having very closed scores are merged together, i.e. only one of them is selected to build the summary.

[1] http://www-nlpir.nist.gov/projects/duc
[2] http://trec.nist.gov/data/qamain.html

6.2 Proposed track at INEX

Since 2008, Question Answering (QA) track at INEX Moriceau et al. (2009) moved into an attempt to bring together Focused Information Retrieval (FIR) intensively experimented in other INEX tracks (previous ad-hoc tracks Geva et al. (2010)) on the one hand, and topic oriented summarization tasks as defined in NIST Text Analysis Conferences (TAC) Dang (2008) on the other hand.

The INEX QA 2009-2010 track Moriceau et al. (2009) aimed to compare the performance of QA, XML/passage retrieval and automatic summarization systems on special XML enriched dumps of the Wikipedia : the 2008 annotated Wikipedia Schenkel et al. (2007) used in the INEX ad-hoc track in 2009 and 2010.

Two types of questions were considered. The first type was factual questions which require a single precise answer to be found in the corpus if it exists. The second type consisted of more complex questions whose answers required a multi-document aggregation of passages with a maximum of 500 words exclusively.

Like for the *2010 ad-hoc restricted focus task*, systems had to make a selection of the most relevant information, the maximal length of the abstract being fixed. Therefore focused IR systems could just return their top ranked passages meanwhile automatic summarization systems need to be combined with a document IR engine. The main difference between the QA long type answer task and the *ad-hoc restricted focus* one is that in QA, readability of answers Pitler et al. (2010) is as important as the informative content. Both need to be evaluated. Therefore answers cannot be any passage of the corpus, but at least well formed sentences. As a consequence, informative content of answers cannot be evaluated using standard IR measures since QA and automatic summarization systems do not try to find all relevant passages, but to select those that could provide a comprehensive answer. Several metrics have been defined and experimented with at DUC Nenkova and Passonneau (2004) and TAC workshops Dang (2008). Among them, Kullback-Leibler (KL) and Jenssen-Shanon (JS) divergences have been used Louis and Nenkova (2009) to evaluate the informativeness of short summaries based on a bunch of highly relevant documents. In this edition we used the KL one to evaluate the informative content of the long answers by comparing their n-gram distributions with those from 4 highly relevant Wikipedia pages.

In 2009 a set of encyclopedic questions about ad-hoc topics was released Moriceau et al. (2009). The idea was that informativeness of answers of encyclopedic questions could be evaluated based on the ad-hoc qrels Geva et al. (2010). This year, a set of

"real" questions from *Over-Blog*[3] website logs not necessarily meant for the Wikipedia was proposed. A state of the art IR engine powered by Indri was also made available to participants. It allowed the participation of seven summarization systems for the first time at INEX. These systems only considered long type answers and have been evaluated on the 2010 subset. Only two standard QA systems participated to the factual question sub-track. Therefore most of QA@INEX 2010 results are about summarization systems versus a state of the art restricted focused IR system.

Like in recent FIR INEX tasks, the corpus is a clean XML extraction of the content of a dump from Wikipedia. However QA track at INEX differs from current FIR and TAC summarization tasks on the evaluation metrics they use to measure both informativeness and readability. Following Louis and Nenkova (2009); Saggion et al. (2010), informativeness measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of relevant passages is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of INEX QA track. By contrast, readability evaluation is completely manual and cannot be reproduced on unofficial runs. It is based on questionnaires pointing out possible syntax problems, broken anaphora, massive redundancy or other major readability problems.

Therefore QA tasks at INEX moved from the usual IR *query / document* paradigm towards *information need / text answer*. More specifically, the task to be performed by the participating groups of INEX 2011 was contextualizing tweets, *i.e.* answering questions of the form "what is this tweet about?". The general process involved:

- Tweet analysis,
- Passage and/or XML element retrieval,
- Construction of the answer.

We target systems efficient on small terminals like smart phones, based on local resources that do not require a network access, gathering non factual contextual information that is scattered around local resources. Off-line applications on portable devices are useful to reduce the network load and safer.

Answers could contain up to 500 words. It has been required that the answer uses only elements previously extracted from the document collection. Answers needed to be a concatenation of textual passages from the Wikipedia dump.

[3]http://www.over-blog.com/

To constitute the pool of RPs, the informativeness of all returned passages for a subset of 50 tweets has been assessed by organizers. The pool of RPs included all passages considered as relevant by at least one assessor (each passage being submitted to two assessors). We regarded as informative passages that both contain relevant information but also contained as little non-relevant information as possible (the result is specific to the question). Long passages including several sentences have often been considered as uninformative because they included too much non relevant information. Furthermore, informativeness of a passage was established exclusively based on its textual content, and not on the documents from which it was extracted. Despite the use of a pool of RPs, the informativeness value of answers did not only rely on the number of its RPs, but also on lexical overlap with other RPs. We found out that evaluating informativeness based on lexical overlap with a pool of RPs is robust if the variety of participant systems is large enough and includes strong baselines.

6.3 Task description

The underlying scenario is to provide the user with synthetic contextual information when receiving a message like a tweet. The task is not to find an exact answer in a database of facts, but to bring out the background of the message exclusively based on its textual content. Therefore the answer needs to be built by aggregation of textual passages grasped from the resource (Wikipedia in our case). For some topics, there can be too many relevant passages that cannot be all inserted in the answer, requiring some summarization process that preserves overall informativeness. For others, only few information can be available and the answer should be shorter than expected pointing out the lack of available information.

In this edition, we have considered a recent dump of the Wikipedia. Since we target non factual answers but short contextualizing texts, we removed all the info boxes and the external references, leaving only the textual content with all its document structure (title, abstract, sections, subtitles and paragraphs) and its internal references (links towards other pages).

We wanted to consider only highly informative tweets. In this attempt to define a contextualizing task, we chose to follow the New York Times (NYT) Twitter account. As soon as the NYT publishes an article on its website, it tweets the title of this article, with its URL. We thus considered these tweets. Therefore the task had become *"given a NYT title, find and summarize all available background information in the Wikipedia"*. We also added the first sentence of the related NYT article as a hint, but only few

runs used this hint and none of the participants reported using NYT paper content: all tried to tackle the contextualization task in an off-line approach using only the available corpus.

The aggregated answers had a maximum of 500 words each and have been evaluated according to:

- Their informativeness (how much they overlap with relevant passages),
- Their readability (assessed by evaluators).

The informativeness of a summary cannot be evaluated without its readability since informative content measures tend to favor syntactic dense summaries. It is often possible to increase an informativeness score by weakening its discursive structure and thus its readability Pitler et al. (2010).

We provided the participants with a state of the art system derived from San-Juan and Ibekwe-Sanjuan (2009); Chen et al. (2010). Participants had to improve its informative performance without weakening too much the readability of its results.

It was initially announced that readability would be evaluated by participants according to the "last point of interest", *i.e.* the first point after which the text becomes unreadable because of:

- syntactic incoherence,
- unsolved anaphora,
- redundancy,
- other problems.

After discussion between organizers and participants at the INEX 2011 workshop, it was finally decided to disclaim considering only the last point of interest because it relied too much on assessors' subjectivity but to mark all readability issues for every sentence in a summary. It was also decided to evaluate the readability independently from the topic to be contextualized and to read all passages, even if redundant. This increased the workload left to participants in readability evaluation but resulted in a much more refined analysis.

Moreover, several on-line resources have been made available to facilitate participation and experiment the metrics. These resources available *via* a unique web interface at *http://termwatch.es/Term2IR* included:

1. a document index powered by Indri,
2. a sentence and Part of Speech tagger powered by the TreeTagger,
3. a summarization and Multi-Word Term extractor powered by TermWatch,
4. a tool for automatic evaluation of summary informativeness powered by FRESA,
5. links to document source on the TopX web interface.

6.4 Document Collection

From 2009 to 2010, QA track at INEX worked on the ad-hoc Wikipedia document collection. In 2009 we considered questions related to ad-hoc topics, and in 2010, real-user, non factual questions from the OverBlog platform[4]. Best performing systems on this task were state of the art automatic summarizers that pick up few Wikipedia pages related to the question and provided a summary as answer.

The document collection has been built based on a dump of the English Wikipedia from April 2011. Since we target a plain XML corpus for an easy extraction of plain text answers, we removed all notes and bibliographic references that are difficult to handle and kept only the 3,217,015 non empty Wikipedia pages (pages having at least one section).

Resulting documents are made of a title (`title`), an abstract (`a`) and sections (`s`). Each section has a sub-title (`h`). Abstract and sections are made of paragraphs (`p`) and each paragraph can have entities (`t`) that refer to other Wikipedia pages.

Therefore the resulting corpus follows this DTD:

```
<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
```

[4] http://www.over-blog.com/

```
<?xml version="1.0" encoding="utf-8"?>
<page>
<ID>2001246</ID>
<title>Alvin Langdon Coburn</title>
<s o="1">
<h>Childhood (1882-1899)</h>
<p o="1">Coburn was born on June 11, 1882, at 134 East Springfield
Street in <t>Boston, Massachusetts</t>, to a middle-class family.
His father, who had established the successful firm of
Coburn &amp;amp; Whitman Shirts, died when he was seven. After that he
was raised solely by his mother, Fannie,  who remained the primary
influence in his early life, even though she remarried when he was
a teenager. In his autobiography, Coburn wrote, &amp;quot;My mother was
a remarkable woman of very strong character who tried to dominate
my life. It was a battle royal all the days of our life
together.&amp;quot;</p>
<p o="2">In 1890 the family visited his maternal uncles in
Los Angeles, and they gave him a 4 x 5 Kodak camera. He immediately
fell in love with the camera, and within a few years he had developed
a remarkable talent for both visual composition and technical
proficiency in the <t>darkroom</t>. (...)</p>
(...)
</page>
```

Figure 6.1: An example of a cleaned Wikipedia XML article.

```
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)><!ELEMENT a (p+)>
<!ELEMENT s (h, p+)>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA | t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>
```

Figure 6.1 shows an example of such a cleaned article. We have released the scripts used to generate this corpus. They process any recent XML dump of the Wikipedia in two steps:

- a light `awk` command to remove in a single pass all external references, info boxes and notes using a fast substring extraction function based on index function (GNU

implementation of strchr C ISO function).

- a `perl` program that generates the XML using regular expressions to detect and encapsulate document structure and internal links. It also works in a single pass.

Once generated, it is necessary to check if the resulting large XML file (between 8 and 12 Gb for recent Wikipedia dumps) is valid. We use the Perl TWIG library by Michel Rodriguez[5] for that. This is a robust library that can process large XML files page by page and fix eventual illformed ones.[6] Current indexers like Indri do not parse such a large XML file and require to split it into pages organized in some folder structure avoiding too large folders. We also made available a Perl program that dispatches Wikipedia pages in 1000 folders. This process can take hours because of numerous file operations.

6.5 Topics

A total set of 155 questions has been made available in 2010:

1. 85 related to 2009 ad-hoc topics,
2. 70 from Over-Blog logs.

Informativeness of answers to questions in first category can be partially evaluated based on qrel from ad-hoc 2009 INEX track meanwhile evaluation in 2010 focused on the second category.

For each question we have selected four highly relevant Wikipedia pages from which we have extracted the most relevant sections. Questions for which there were too few relevant passages were not submitted to participants. These passages that were not publicly available have been then used as reference text to evaluate long type answers using KL divergence.

[5] http://search.cpan.org/ mirod/

[6] We had to manually correct few errors on the April 2011 Wikipedia dump due to encoding errors in the original dump file itself, but we did not have error anymore in the last Wikipedia dump from November 2011. For the 2011 INEX edition, we used the corrected April 2011 dump.

In 2011, the QA track started experimenting tweets instead of real questions. There the overlap between topics and Wikipedia content becomes much weaker than previously. It was thus decided to move to a more recent and simplified dump of Wikipedia. The new corpus was made available in October 2011 leaving two months to participants for their experiments. This corpus generation process has been completely automatized and can be apply to any XML Wikipedia dump.

The topic data set was composed of 132 tweets by the NYT released on the July 20th 2011 and having a URL towards the NYT website. Each topic includes the tweet which is often the title of an article just released and the first sentence of the related article. An example is provided below:

```
<topic id="2011005">
  <title>Heat Wave Moves Into Eastern U.S</title>
  <txt>The wave of intense heat that has enveloped much of the
    central part of the country for the past couple of weeks is
    moving east and temperatures are expected to top the 100-degree
    mark with hot, sticky weather Thursday in cities from
    Washington, D.C., to Charlotte, N.C.</txt>
</topic>
```

All these topics were twitted three months after the Wikipedia dump used to build the corpus, therefore we had to manually check if there was any related information in the document collection[7]

6.6 Submission requirements

Participants could submit up to three runs. Despite the fact that manual runs were allowed if there was at least one automatic, all submitted official runs have been registered as fully automatic.

Results were lists of passages extracted from the corpus. Two non consecutive passages had to be presented separately. Results in a single run could not include more

[7] The resulting 132 topics come from an initial set of 205 tweets after removing duplicates due to single subjects producing several papers (like different testimonies and opinion papers about the same subject) and only few tweets for which there was no overlap with the Wikipedia. Hence, the 132 selected topics represent more than 64% of the tweets released by the NYT in one day.

than 500 words per topic. Any string of alphanumeric characters outside XML tags, without space or punctuation, was considered as a single word.

The format for results was a variant of the familiar TREC format with additional fields:[8]

```
<qid> Q0 <file> <rank> <rsv> <run_id> <column_7> <column_8>
```

where:

- The first column `qid` is the topic number.
- The second column is currently unused and should always be `Q0`. It is just a formating requirement used by the evaluation programs to distinguish between official submitted runs and q-rels.
- The third column `file` is the file name (without `.xml`) from which a result is retrieved, which is identical to the `<id>` of the Wikipedia document. It is only used to retrieve the raw text content of the passage, not to compute document retrieval capabilities. In particular, if two results only differ by their document id (because the text is repeated in both), then they will be considered as identical and thus redundant.
- The fourth column `rank` indicates the order in which passages should be read for readability evaluation, this differs from the expected informativeness of the passage who is indicated by the score `rsv` in the fifth column. Therefore, these two columns are not necessarily correlated. Passages with highest scores in the fifth column can be scattered at any rank in the result list for each topic.
- The sixth column `run_id` is called the "run tag" and should be a unique identifier for the participant group and for the method used.
- The remaining two columns indicate the selected passage in the document mentioned in the third field. Participants could refer to these passages as File Offset Lengths (FOL) like in usual INEX FIR tasks or directly give the raw textual content of the passage. However, computing character offsets can be tricky dependent on the text encoding and Wikipedia often mixes different encodings. Therefore all

[8] The XML format to submit results originally proposed in 2010 was dismissed since it was never used by participants because of its useless extra complexity. However if the task evolves in the following years towards more complex results, TREC-like formats will not be sufficient and some XML formating will be required.

participants to this edition chose the alternative raw text format. In this format, each result passage is given as raw text without XML tags and without formatting characters. The only requirement is that the resulting word sequence has to appear at least once in the file indicated in the third field.

Here is an example of such an output:

```
2011001 Q0 3005204 1 0.9999 I10UniXRun1 The Alfred Noble Prize is ...
2011001 Q0 3005204 2 0.9998 I10UniXRun1 The prize was established in ...
2011001 Q0 3005204 3 0.9997 I10UniXRun1 It has no connection to the ...
```

6.7 Evaluation Metrics

2010 Edition

Long type questions require long answers up to 500 words that must be self-contained summaries made of passages extracted from the INEX 2009 corpus. Are considered as words any sequence of letters and digits. An example of a long type question is (#196): *What sort of health benefit has olive oil?* There can be questions of both short and long types, for example a question like *Who was Alfred Nobel?* can be answered by "a chemist" or by a short biography. However, most of the selected long type questions are not associated with obvious name entities and require at least one sentence to be answered.

The informative content of the long type answers were evaluated by comparing the several n-gram distributions in participant extracts and in a set of relevant passages selected manually by organizers. We followed the experiment in Louis and Nenkova (2009) done on TAC 2008 automatic summarization evaluation data. This allows to evaluate directly summaries based on a selection of relevant passages.

Given a set R of relevant passages and a text T, let us denote by $p_X(w)$ the probability of finding an n-gram w from the Wikipedia in $X \in \{R, T\}$. We use standard Dirichlet smoothing with default $\mu = 2500$ to estimate these probabilities over the whole corpus. Word distributions are usually compared using one of these functions:

- Kullback Leibler (KL) divergence:

$$KL(p_T, p_R) = \sum_{w \in R \cup T} p_T(w) \times \log_2 \frac{p_T(w)}{p_R(w)}$$

- Jensen Shannon (JS) divergence:

$$JS(p_T, p_R) = \frac{1}{2}(KL(p_T, p_{T \cup R}) + KL(p_R, p_{T \cup R}))$$

In Louis and Nenkova (2009), the metric that obtained best correlation scores with ROUGE semi-automatic evaluations of abstracts used in DUC and TAC was JS. However, we have observed that JS is too sensitive to abstract size; therefore we finally used KL divergence to evaluate informative content reference texts or passages.

We used the FRESA package[9] to compute both KL and JS divergences between n-grams ($1 \leq n \leq 4$). This package also allows to consider skip n-grams.

Evaluating informative content without evaluating readability does not make sense. It clearly appears that if readability is not considered then the best summarizer would be the random summarizer on n-grams which certainly minimizes KL divergence but produces incomprehensible texts.

In 2010, the readability and coherence are evaluated according to "the last point of interest" in the answer which is the counterpart of the "best entry point" in INEX ad-hoc task. It requires a human evaluation by organizers and participants where the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages.

2011 edition

Systems had to make a selection of the most relevant information, the maximal length of the abstract being fixed. Focused IR systems could just return their top ranked passages meanwhile automatic summarization systems need to be combined with a document IR engine. Both need to be evaluated. Therefore answers cannot be any passage of the corpus, but at least well formed sentences. As a consequence, informative content of answers cannot be evaluated using standard IR measures since QA and automatic summarization systems do not try to find all relevant passages but to select those that

[9] http://lia.univ-avignon.fr/fileadmin/axes/TALNE/Ressources.html

could provide a comprehensive answer. Several metrics have been defined and experimented with at DUC Nenkova and Passonneau (2004) and TAC workshops Dang (2008). Among them, Kullback-Leibler (KL) and Jenssen-Shanon (JS) divergences have been used Louis and Nenkova (2009); Saggion et al. (2010) to evaluate the informativeness of short summaries based on a bunch of highly relevant documents.

In 2010 we intended to use the KL one with Dirichlet smoothing, like in the 2010 Edition SanJuan et al. (2010), to evaluate the informative content of answers by comparing their n-gram distributions with those from all assessed relevant passages. However, in 2010, references were made of complete Wikipedia pages, therefore the textual content was much longer than summaries and smoothing did not introduce too much noise.

This is not the case with the 2011 assessments. For some topics, the amount of relevant passages is very low, less than the maximal summary length. Therefore using any probabilistic metric requiring some smoothing produced very unstable rankings. We thus simply considered absolute log-diff between frequencies. Let T be the set of terms in the reference. For every $t \in T$, we denote by $f_T(t)$ its frequency in the reference and by $f_S(t)$ its frequency in the summary. Adapting the FRESA package available to participants, we computed the divergence between reference and summaries as:

$$Div(T,S) = \sum_{t \in T} |\log(\frac{f_T(t)}{f_T} + 1) - \log(\frac{f_S(t)}{500} + 1)| \quad (6.3)$$

As T we considered three different sets based on the FRESA sentence segmentation, stop word list and lemmatizer:

- Unigrams made of single lemmas (after removing stop-words).
- Bigrams made of pairs of consecutive lemmas (in the same sentence).
- Bigrams with 2-gaps also made of pairs of consecutive lemmas but allowing the insertion between them of a maximum of two lemmas.

As in 2010, bigrams with 2-gaps appeared to be the most robust metric. Sentences are not considered as simple bag of words and it is less sensitive to sentence segmentation than simple bi-grams. This is why bigrams with 2-gaps is our official ranking metric for informativeness.

For readability evaluation, each participant had to evaluate readability for a pool of around 50 summaries of a maximum of 500 words each on an online web interface.

Each summary consisted in a set of passages and for each passage, assessors had to tick four kinds of check boxes. The guideline was the following:

- *Syntax* (S): tick the box if the passage contains a syntactic problem (bad segmentation for example),
- *Anaphora* (A): tick the box if the passage contains an unsolved anaphora,
- *Redundancy* (R): tick the box if the passage contains a redundant information, i.e. an information that has already been given in a previous passage,
- *Trash* (T): tick the box if the passage does not make any sense in its context (*i.e.* after reading the previous passages). These passages must then be considered at trashed, and readability of following passages must be assessed as if these passages were not present.
- If the summary is so bad that you stop reading the text before the end, tick all trash boxes until the last passage.

The assessors did not know the topic corresponding to the summary, and were not supposed to judge the relevance of the text. Only readability was evaluated.

To evaluate summary readability, we consider the number of words (up to 500) in valid passages. We used two metrics based on this:

- **Relaxed metric**: a passage is considered as valid if the T box has not been ticked,
- **Strict metric**: a passage is considered as valid if no box has been ticked.

In both cases, participant runs are ranked according to the average, normalized number of words in valid passages.

6.8 A baseline restricted focused system

The system allows to test on the INEX 2009 ad-hoc corpus the combination of a simple IR passage retrieval system (Indri Language Model) with a baseline summarization system (a fast approximation of Lexrank).

Different outputs are available. The default is a selection of relevant sentences with a link towards the source document in TopX. Sentences have been selected following approximated LexRank scores among the 20 top ranked passages returned by Indri using a Language Model over INEX 2008 corpus. Multiword terms extracted by shallow parsing are also highlighted.

A second possible output gives a baseline summary with less than 500 words, made of the top ranked sentences. The Kullback-Leibler divergence between distributions of n-grams in the summary and in the passages retrieved by Indri are also shown. They are computed using the FRESA package. It is also possible to test any summary against this baseline.

Finally, the passages retrieved by Indri are available, in several formats: raw results in native INEX XML format, raw text, POS tagged text with TreeTagger.

Questions and queries can be submitted in plain text or in Indri language. The following XML tags have been indexed and can be used in the query: *b, bdy, category, causal_agent, country, entry, group, image, it, list, location, p, person, physical_entity, sec, software, table, title*. These are examples of correct queries:

- Who is Charlie in the chocolate factory?
- #1(Miles davis) #1(Charles Mingus) collaboration
- #1(Charles Mingus).p, #combine[p](Charles Mingus)

Let us first give some details on this restricted focus system.

As stated before it starts by retrieving n documents using an Indri language model. These sentences are then segmented into sentences using shallow parsing. Finally sentences are ranked using a fast approximation of LexRank. Basically, we only consider sentences that are at distance two from the query in the intersection graph of sentences. These are sentences that share at least one term with the query, or with another sentence that shares it. The selected sentences are then ranked by entropy.

We evaluated this baseline system on the Ad-hoc restricted focused task, by setting $n = 100$. We then retrieve for each sentence all passages in which the same word sequence appears, with possible insertions. We return the first 1000 characters.

The precision/recall function of this system starts high compared to other participant runs. It gets among automatic runs, the third char precision (0.3434) and the best

iP[0.01] with a value of 0.15 (0.1822 for the best manual run).

The baseline XML-element retrieval/summarization system has been made available for INEX participants. It relies on:

- An index powered by Indri[10] that covers all words (no stop list, Krowetz stemming) and all XML tags.
- A PartOfSpeech tagger powered by TreeTagger[11].
- A fast summarizer algorithm powered by TermWatch[12] introduced in Chen et al. (2010).

The Indri index allows to experiment different types of queries to seek for all passages in the Wikipedia involving terms in the topic. Queries can be usual bag of words, sets of weighted multi-word phrases or more complex structured queries using Indri Language Metzler and Croft (2004). All extracted passages are segmented into sentences and PoS tagged using the TreeTagger. Sentences are then scored using TermWatch based on their *nominals* (i.e. its nouns and adjectives). Let Φ be the set of sentences. If for each sentence $\phi \in \Phi$, we denote by φ_ϕ the set of its nominals, then the sentence score Θ_ϕ computed in Chen et al. (2010) is:

$$\Theta_\phi = \sum_{\varphi_\phi \cap \varphi_\tau \neq \emptyset}^{\tau \in \Phi} \sum_{\varphi_\tau \cap \varphi_\sigma \neq \emptyset}^{\sigma \in \Phi} |\varphi_\phi \cap \varphi_\tau| \times |\varphi_\tau \cap \varphi_\sigma| \tag{6.4}$$

The idea is to weight the sentences according to the number of sentences in their neighborhood (sentences sharing at least one nominal). This gives a fast approximation of TextRank or LexRank scores Chen et al. (2010). Sentences are then ranked by decreasing score, only the top ranked are used for a summary of less than 500 words. The selected sentences are then re-ordered following the Indri score of the passage from which they have been extracted and the order of the sentences in these passages. This baseline summary can be computed on the fly, generating the summary taking less time than processing the query by Indri.

This system has been made available online to participants through a web interface[13]. A Perl API running on Linux to query the server was also released. By default, this API takes as input a tabulated file with three fields: topic names, selected output

[10] http://www.lemurproject.org/
[11] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[12] http://data.termwatch.es
[13] http://qa.termwatch.es

format and query. The output format can be the baseline summary or the first 50 retrieved documents in raw text, PoS tagged or XML source. An example of such a file allowing to retrieve 50 documents per topic based on their title was also released.

The web interface also allows to evaluate the resulting summary or user's one against the retrieved documents using Kullback-Leibler (KL) measure. This content summary evaluation also gives a lower bound using a random set of 500 words extracted from the texts and an upper bound using an empty summary. Random summaries naturally reach the closest word distributions but they are clearly unreadable.

In 2010 and 2011, two baselines were then computed using the approach described in Chen et al. (2010) and added to the pool of official submissions:

- Baseline_sum using only topic titles as bag of word queries and top ranked 50 full documents retrieved by Indri to build the summary.
- Baseline_mwt using the same process but returning only the Noun Phrases in the selected sentences to simulate a baseline run for Automatic Terminology Extractors.

6.9 Results

In this task, readability of answers Pitler et al. (2010) is as important as the informative content. Summaries must be easy to read as well as relevant. These two properties have been evaluated separately by two distinct measures: *informativeness* and *readability*.

6.9.1 General comments

In 2011 23 valid runs by 11 teams from 6 countries (Brasil, Canada, France, India, Mexico, Spain) were submitted. All runs are in raw text format and almost all participants used their own summarization system. Only three participants did not use the online Indri IR engine. Some participants used the Perl API to query the Indri Index with expanded queries based on semantical resources. Only one participant used XML tags.

The total number of submitted passages is 37,303. The median number of distinct passages per topic is 284.5 and the median length in words is 26.9. This relative small

amount of distinct passages could be due to the fact that most of the participants used the provided Indri index with its Perl API.

In 2010, for each question we have selected four highly relevant Wikipedia pages from which we have extracted the most relevant sections. Questions for which there were too few relevant passages were not submitted to participants. These passages that were not publicly available have been then used as reference text to evaluate long type answers using KL divergence.

In 2011, Informativeness evaluation has been performed by organizers on a pool of 50 topics. For each of these topics, all passages submitted have been evaluated. Only passages starting and ending by the same 25 characters have been considered as duplicated, therefore short sub-passages could appear twice in longer ones. For each topic, all passages from all participants have been merged and displayed to the assessor in alphabetical order. Therefore, each passage informativeness has been evaluated independently from others, even in the same summary. The structure and readability of the summary was not assessed in this specific part, and assessors only had to provide a binary judgment on whether the passage was worth appearing in a summary on the topic, or not. This approach handicaps runs based on short passages extracted from the Wikipedia, since very short passages can be difficult to assess on their own and tend not to be included in the pool of relevant passages.

To check that the resulting pool of relevant answers is sound, a second automatic evaluation for informativeness of summaries has been carried out with respect to a reference made of the NYT article corresponding to the topic. Official evaluation could not be based on these references since most of these articles were still available on the NYT website or could have been used by participants who are NYT readers. Nevertheless, a strong correlation between the ranking based on the assessed pool of relevant passages and the one based on NYT articles would be an indication of assessment soundness.

In 2010, we received runs for long type questions from seven participants. All of these participants generate summaries by sentence extraction. This helps readability even if it does not ensure general coherence. Extracts made of long sentences without anaphora are often more coherent but have higher KL scores. To retrieve documents, all participants used the IR engine powered by Indri, available at track resources webpage[14].

As expected, baseline-restricted focused IR system minimizes KL divergence but the resulting readability is poor. Meanwhile the system having best readability favors long sentences and gets highest divergence figures. The most sophisticated summary ap-

[14]http://qa.termwatch.es/

proach is the Cortex system (860) which reaches a compromise between KL divergence and readability.

But query formulation to retrieve documents looks also important, the approach based on query enrichment with related MultiWord Terms automatically extracted from top ranked documents, gets similar divergence scores. Meanwhile this is a system slightly adapted from the focused IR system used in previous INEX 2008 and 2009 ad-hoc track SanJuan and Ibekwe-Sanjuan (2009); Ibekwe-Sanjuan and SanJuan (2008).

Surprisingly sentence JS minimization does not seem to minimize overall KL divergence. This system ranks sentences in decreasing order according to their JS divergence with the query and the retrieved documents.

Only score differences between the baseline and the other systems were significant.

The standard deviation among systems KL divergences varies. The ten question minimizing standard deviation and, therefore, getting most similar answers among systems are:

2010044 What happened to the president of Rwanda death?

2010107 What are the symptoms of a tick bite?

2010096 How to make rebellious teenager obey you?

2010066 How much sadness is normal breakup?

2010062 How much is a typical sushi meal in japan?

2010083 What are the Refugee Camps in DRC?

2010046 How to get Aljazera sports?

2010047 How to be a chef consultant?

2010005 Why did Ronaldinho signed for Barcelona?

2010049 Where can I find gold sequined Christain Louboutin shoes?

All these questions contain at least one named entity that refers to a Wikipedia page. Therefore, the systems mostly built their answer based on the textual content of this page and KL divergence is not accurate enough to discriminate among them.

On the contrary, the 10 following questions are the top ten that maximized standard deviation and have the greatest impact in the ranking of the systems:

2010093 Why is strategy so important today?

2010114 What is epigenetics and how does it affect the DNA/genes in all of our cells?

2010009 What does ruddy complexion mean?

2010066 What do nice breasts look like?

2010022 How to get over soul shock?

2010092 How to have better sex with your partner?

2010080 How to be physically attractive and classy?

2010014 Why is it so difficult to move an mpeg into imovie?

2010010 What do male plants look like?

2010075 WHAT IS A DUAL XD ENGINE?

Clearly, these questions are not encyclopedic ones and do not refer to particular Wikipedia pages. Meanwhile partial answers exist in the Wikipedia but they are spread among several articles.

In 2011, all passages within a consistent pool of 50 topics were thoroughly evaluated by organizers. This represents 14,654 passages, among which 2,801 have been judged as relevant.

This assessment was intended to be quite generous towards passages. All passages concerning a protagonist of the topic are considered relevant, even if the main subject of the topic is not addressed. The reason is that missing words in the reference can lead to artificial increase of the *divergence*, which is a known and not desirable side effect of this measure.

Coming to readability, A total of 1,310 summaries, 28,513 passages from 53 topics have been assessed. All participants succeeded in evaluating more than 80% of the assigned summaries. The resulting 53 topics include all of those used for informativeness assessment.

None of the submitted participant runs outperformed Baselinesum (Baseline with complete summaries).

The other baseline restricted to Multi Word Noun Phrases was considered as unreadable by most assessors except by one who is a specialist in terminology and considered as acceptable any NP that corresponds to a real Multi Word Term.

6.9.2 Baseline

For 2010 Ad-hoc restricted focused task, we retrieve and return for each sentence any passage in which the same word sequence appears, with possible insertions. The precision/recall function starts very high for this system compared to other participant runs, but then drops very quickly.

Results are presented in Table 6.1.

Table 6.1: Focused retrieval results on the Restricted Focused task in terms of Mean Average Precision (MAP).

Institute	**Runs**	MAP
University Pierre et Marie Curie - LIP6	`LIP6-OWPCparentFo`	0.4125
Doshisha University	`DURF10SIXF`	0.3884
LIA - University of Avignon	**`LIAenertexTopic`**	**0.3434**
Peking University	`40p167`	0.3370

In the 2011 QA Task, all systems above the baseline combine a full document retrieval engine with a summarization algorithm. The three top ranked runs, all by IRIT, did not use the API provided to participants meanwhile all other runs improving the baseline used it only to query the Indri Index, some applying special query expansion techniques. None of the participants used this year the baseline summarization system which ranks 7th among all runs when returning full sentences (`Baselinesum`) and 19th when returning only noun phrases (`Baselinemwt`).

Dissimilarity values are very closed, however differences are often statistically significant as shown in table 6.3. In particular, top four runs are significantly better than all others. It seems that these runs carried out specific NLP post-processing. It also appears that almost all runs above Baselinesum are significantly better than those under the same baseline, meanwhile differences among runs ranked between the two baselines are rarely significant.

Rank	Run	unigram	bigram	**with 2-gap**	Average
1	ID12_IRIT_default	0.0486	0.0787	**0.1055**	0.0787
2	ID12_IRIT_07_2_07_1_dice	0.0488	0.0789	**0.1057**	0.0789
3	ID12_IRIT_05_2_07_1_jac	0.0491	0.0792	**0.1062**	0.0793
4	ID129_Run1	0.0503	0.0807	**0.1078**	0.0807
5	ID129_Run2	0.0518	0.0830	**0.1106**	0.0830
6	ID128_Run2	0.0524	0.0834	**0.1110**	0.0834
7	ID138_Run1	0.0524	0.0837	**0.1115**	0.0837
8	ID18_Run1	0.0526	0.0838	**0.1117**	0.0839
9	ID126_Run1	0.0535	0.0848	**0.1125**	0.0848
10	Baselinesum	0.0537	0.0859	**0.1143**	0.0859
11	ID126_Run2	0.0546	0.0863	**0.1144**	0.0863
12	ID128_Run3	0.0549	0.0869	**0.1151**	0.0868
13	ID129_Run3	0.0549	0.0869	**0.1152**	0.0869
14	ID46_JU_CSE_run1	0.0561	0.0877	**0.1156**	0.0876
15	ID46_JU_CSE_run2	0.0561	0.0877	**0.1156**	0.0876
16	ID62_Run3	0.0565	0.0887	**0.1172**	0.0887
17	ID123_I10UniXRun2	0.0561	0.0885	**0.1172**	0.0885
18	ID128_Run1	0.0566	0.0889	**0.1174**	0.0889
19	Baselinemwt	0.0558	0.0886	**0.1179**	0.0887
20	ID62_Run1	0.0566	0.0892	**0.1180**	0.0892
21	ID123_I10UniXRun1	0.0567	0.0895	**0.1183**	0.0894
22	ID62_Run2	0.0572	0.0900	**0.1188**	0.0899
23	ID124_UNAMiiR12	0.0607	0.0934	**0.1221**	0.0933
24	ID123_I10UniXRun3	0.0611	0.0946	**0.1239**	0.0945
25	ID124_UNAMiiR3	0.0628	0.0957	**0.1248**	0.0957

Table 6.2: Informativeness results from manual evaluation using equation 6.3 (official results are "with 2-gap").

	ID12_IRIT_default	ID12_IRIT_07_2_07_1_dice	ID12_IRIT_05_2_07_1_jac	ID129_Run1	ID129_Run2	ID128_Run2	ID138_Run1	ID18_Run1	ID126_Run1	Baselinesum	ID126_Run2	ID128_Run3	ID129_Run3	ID46_JU_CSE_run1	ID46_JU_CSE_run2	ID62_Run3	ID123_I10UniXRun2	ID128_Run1	Baselinemwt	ID62_Run1	ID123_I10UniXRun1	ID62_Run2	ID124_UNAMiiR12	ID123_I10UniXRun3	ID124_UNAMiiR3
ID12_IRIT_default	-	-	1	-	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_07_2_07_1_dice	-	-	1	-	1	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_05_2_07_1_jac	1	1	-	-	1	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run1	-	-	-	-	2	1	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run2	2	1	1	2	-	-	-	-	-	3	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID128_Run2	2	2	2	1	-	-	-	-	-	1	2	3	2	2	2	3	3	3	3	3	3	3	3	3	3
ID138_Run1	2	2	2	3	-	-	-	-	-	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID18_Run1	3	2	2	2	-	-	-	-	-	-	-	-	1	1	1	3	3	3	3	3	3	3	3	3	3
ID126_Run1	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	2	2	3	3	3	3	3	3	3	3
Baselinesum	3	3	3	3	3	1	1	-	-	-	-	-	-	-	-	-	2	1	3	2	2	3	3	3	3
ID126_Run2	3	3	3	3	2	2	2	-	-	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3
ID128_Run3	3	3	3	3	2	3	2	-	-	-	-	-	-	-	-	-	-	1	1	1	1	2	3	3	3
ID129_Run3	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	1	1	1	2	3	3	3
ID46_JU_CSE_run1	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	3	3
ID46_JU_CSE_run2	3	3	3	3	2	2	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	3	3
ID62_Run3	3	3	3	3	3	3	3	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	2	3	3
ID123_I10UniXRun2	3	3	3	3	3	3	3	3	2	2	1	-	-	-	-	-	-	-	-	-	-	1	3	3	3
ID128_Run1	3	3	3	3	3	3	3	3	3	1	2	1	-	-	-	-	-	-	-	-	-	-	2	3	3
Baselinemwt	3	3	3	3	3	3	3	3	3	3	2	1	1	-	-	-	-	-	-	-	-	-	3	3	3
ID62_Run1	3	3	3	3	3	3	3	3	3	2	2	1	1	-	-	-	-	-	-	-	-	-	2	3	3
ID123_I10UniXRun1	3	3	3	3	3	3	3	3	3	2	2	1	1	-	-	-	-	-	-	-	-	-	2	3	3
ID62_Run2	3	3	3	3	3	3	3	3	3	3	3	2	2	-	-	-	1	-	-	-	-	-	1	3	3
ID124_UNAMiiR12	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	2	3	2	2	1	-	-	3
ID123_I10UniXRun3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-	-
ID124_UNAMiiR3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-

Table 6.3: Statistical significance for official results in table 6.2 (t-test, $1 : 90\%$, $2 = 95\%$, $3 = 99\%$, $\alpha = 5\%$).

To check that this reference was not biased, the same 50 topics have been also automatically evaluated against the corresponding NYT article, *i.e.* taking as reference the article published under the tweeted title. None of the participants reported having used this content even though part of it was publicly available on the web.

Results are presented in Table 6.4. It appears that correlation between the two rankings is quite high (Kendall's $\tau = 0.67$, Pearson's product-moment correlation = 88%, p-value $< 9.283e^{-9}$) suggesting that our approach of selecting reference text from a pool of participant runs plus the baselines is sufficient.

All previous evaluations have been carry out using FRESA package which includes a special lemmatizer. We provided the participants with a standalone evaluation toolkit based on Porter stemmer. Based on participant feedback after the release of the official results, we introduced in this package a normalized ad-hoc dissimilarity defined as following using the same notations as in equation 6.3:

$$Dis(T,S) = \sum_{t \in T} \frac{f_T(t)}{f_T} \times \left(1 - \frac{\min(\log(P), \log(Q))}{\max(\log(P), \log(Q))}\right) \quad (6.5)$$

$$P = \frac{f_T(t)}{f_T} + 1 \quad (6.6)$$

$$Q = \frac{f_S(t)}{f_S} + 1 \quad (6.7)$$

The idea is to have a dissimilarity which complement has similar properties to usual IR Interpolate Precision measures. Actually, $1 - Dis(T,S)$ increases with the Interpolated Precision at 500 tokens where Precision is defined as the number of word n-grams in the reference. The introduction of the log is necessary to deal with highly frequent words.

Table 6.5 shows results using this evaluation toolkit implementing basic stemming and normalized dissimilarity 6.5. Again, the correlation with official results in Table 6.2 is quite high (Kendall's τ =89%, Pearson's product-moment correlation = 96%, p-value $< 4e^{-11}$).

This normalized metric does not allow to distinguish between top ranked runs above the baseline as shown by statistical significance tests reported in table 6.6 but it does among runs between the two baselines.

Conserning readbility, in 2011, results are presented in Table 6.7.

The high score of the baseline can be explained by the fact that formula 6.4 favors sentences with numerous Multi Word Noun Phrases. These particular sentences tend to be long, with few pronouns, thus few broken anaphora. The drawback of this Baseline

Rank	Run	unigram	bigram	**with 2-gap**	Average
1	ID12_IRIT_05_2_07_1_jac	0.0447	0.0766	**0.1049**	0.0766
2	ID12_IRIT_07_2_07_1_dice	0.0447	0.0767	**0.1049**	0.0766
3	ID12_IRIT_default	0.0447	0.0767	**0.1049**	0.0767
4	ID129_Run1	0.0456	0.0777	**0.1060**	0.0777
5	ID18_Run1	0.0462	0.0779	**0.1061**	0.0779
6	Baselinesum	0.0460	0.0781	**0.1065**	0.0781
7	ID126_Run1	0.0460	0.0781	**0.1065**	0.0781
8	ID128_Run2	0.0461	0.0782	**0.1066**	0.0782
9	ID138_Run1	0.0461	0.0782	**0.1066**	0.0782
10	ID129_Run2	0.0468	0.0788	**0.1071**	0.0787
11	ID129_Run3	0.0468	0.0789	**0.1072**	0.0788
12	ID126_Run2	0.0469	0.0789	**0.1073**	0.0789
13	ID128_Run3	0.0469	0.0789	**0.1073**	0.0789
14	ID123_I10UniXRun1	0.0471	0.0791	**0.1075**	0.0791
15	Baselinemwt	0.0475	0.0794	**0.1077**	0.0794
16	ID62_Run1	0.0473	0.0793	**0.1077**	0.0793
17	ID128_Run1	0.0475	0.0795	**0.1079**	0.0795
18	ID62_Run3	0.0476	0.0796	**0.1080**	0.0796
19	ID62_Run2	0.0477	0.0797	**0.1080**	0.0797
20	ID123_I10UniXRun2	0.0477	0.0797	**0.1080**	0.0797
21	ID123_I10UniXRun3	0.0483	0.0804	**0.1087**	0.0803
22	ID46_JU_CSE_run1	0.0487	0.0807	**0.1089**	0.0806
23	ID46_JU_CSE_run2	0.0487	0.0807	**0.1090**	0.0807
24	ID124_UNAMiiR12	0.0493	0.0812	**0.1094**	0.0812
25	ID124_UNAMiiR3	0.0505	0.0823	**0.1104**	0.0823

Table 6.4: Informativeness results automatic evaluation against NYT article using equation 6.3.

Rank	Run	unigram	bigram	**with 2-gap**
1	ID12_IRIT_default	0.8271	0.9012	**0.9028**
2	ID126_Run1	0.7982	0.9031	**0.9037**
3	ID12_IRIT_07_2_07_1_dice	0.8299	0.9032	**0.9053**
4	ID129_Run1	0.8167	0.9058	**0.9062**
5	ID12_IRIT_05_2_07_1_jac	0.8317	0.9046	**0.9066**
6	ID128_Run2	0.8034	0.9091	**0.9094**
7	ID138_Run1	0.8089	0.9150	**0.9147**
8	ID129_Run2	0.8497	0.9252	**0.9253**
9	ID126_Run2	0.8288	0.9306	**0.9313**
10	ID128_Run3	0.8207	0.9342	**0.9350**
11	Baselinesum	0.8363	0.9350	**0.9362**
12	ID18_Run1	0.8642	0.9368	**0.9386**
13	ID129_Run3	0.8563	0.9436	**0.9441**
14	ID46_JU_CSE1	0.8807	0.9453	**0.9448**
15	ID46_JU_CSE2	0.8807	0.9452	**0.9448**
16	ID128_Run1	0.8379	0.9492	**0.9498**
17	ID62_Run3	0.8763	0.9588	**0.9620**
18	ID123_I10UniXRun2	0.8730	0.9613	**0.9640**
19	ID62_Run1	0.8767	0.9667	**0.9693**
20	ID62_Run2	0.8855	0.9700	**0.9723**
21	ID123_I10UniXRun1	0.8840	0.9699	**0.9724**
22	ID124_UNAMiiR12	0.9286	0.9729	**0.9740**
23	Baselinemwt	0.9064	0.9777	**0.9875**
24	ID124_UNAMiiR3	0.9601	0.9896	**0.9907**
25	ID123_I10UniXRun3	0.9201	0.9913	**0.9925**

Table 6.5: Informativeness results from manual evaluation 6.5

	ID12_IRIT_default	ID126_Run1	ID12_IRIT_07_2_07_1_dice	ID129_Run1	ID12_IRIT_05_2_07_1_jac	ID128_Run2	ID138_Run1	ID129_Run2	ID126_Run2	ID128_Run3	Baseline_sum	ID18_Run1	ID129_Run3	ID46_JU_CSE_run2	ID46_JU_CSE_run1	ID128_Run1	ID62_Run3	ID123_I10UniXRun2	ID62_Run1	ID62_Run2	ID123_I10UniXRun1	ID124_UNAMiiR12	Baseline_mwt	ID124_UNAMiiR3	ID123_I10UniXRun3
ID12_IRIT_default	-	-	-	-	1	-	-	-	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
ID126_Run1	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_07_2_07_1_dice	-	-	-	-	-	-	-	-	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID129_Run1	-	-	-	-	-	-	-	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID12_IRIT_05_2_07_1_jac	1	-	-	-	-	-	-	-	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3
ID128_Run2	-	-	-	-	-	-	-	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
ID138_Run1	-	-	-	-	-	-	-	-	2	1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
ID129_Run2	-	1	-	2	-	1	-	-	-	-	-	-	1	2	2	2	3	3	3	3	3	3	3	3	3
ID126_Run2	2	2	1	2	1	2	2	-	-	-	-	-	-	-	-	2	2	3	3	3	3	3	3	3	3
ID128_Run3	2	2	1	2	1	2	1	-	-	-	-	-	-	-	-	1	2	3	3	3	3	3	3	3	3
Baseline_sum	2	2	2	3	1	2	2	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3
ID18_Run1	2	2	2	3	2	2	3	-	-	-	-	-	-	-	-	-	2	3	3	3	3	3	3	3	3
ID129_Run3	2	3	2	3	2	3	3	1	-	-	-	-	-	-	-	-	1	2	3	3	3	3	3	3	3
ID46_JU_CSE_run2	3	3	2	3	2	3	3	2	-	-	-	-	-	-	-	-	1	2	3	3	3	2	3	3	3
ID46_JU_CSE_run1	3	3	2	3	2	3	3	2	-	-	-	-	-	-	-	-	1	2	3	3	3	2	3	3	3
ID128_Run1	3	3	3	3	3	3	3	2	2	1	-	-	-	-	-	-	-	2	3	3	3	2	3	3	3
ID62_Run3	3	3	3	3	3	3	3	3	2	2	2	2	1	1	1	-	-	-	-	-	-	-	3	3	3
ID123_I10UniXRun2	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	2	-	-	1	2	2	-	3	3	3
ID62_Run1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	1	-	-	-	-	3	3	3
ID62_Run2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	2	-	-	-	-	2	3	3
ID123_I10UniXRun1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	2	-	-	-	-	2	3	3
ID124_UNAMiiR12	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	-	-	-	-	-	-	1	3	2
Baseline_mwt	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	-	-	-
ID124_UNAMiiR3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	-	-	-
ID123_I10UniXRun3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	-	-	-

Table 6.6: Statistical significance for manual evaluation 6.5 (t-test, 1 : 90%, 2 = 95%, 3 = 99%, $\alpha = 5\%$).

Relaxed metric			**Strict metric**		
Rank	Run id	Score	Rank	Run id	Score
1	Baseline_sum	447.3019	1	Baseline_sum	409.9434
2	ID46_JU_CSE_run1	432.2000	2	ID129_Run1	359.0769
3	ID128_Run2	417.8113	3	ID129_Run2	351.8113
4	ID12_IRIT_default	417.3462	4	ID126_Run1	350.6981
5	ID46_JU_CSE_run2	416.5294	5	ID46_JU_CSE_run1	347.9200
6	ID129_Run1	413.6604	6	ID12_IRIT_05_2_07_1_jac	344.1154
7	ID129_Run2	410.7547	7	ID12_IRIT_default	339.9231
8	ID12_IRIT_05_2_07_1_jac	409.4038	8	ID12_IRIT_07_2_07_1_dice	338.7547
9	ID12_IRIT_07_2_07_1_dice	406.3962	9	ID128_Run2	330.2830
10	ID126_Run1	404.4340	10	ID46_JU_CSE_run2	330.1400
11	ID138_Run1	399.3529	11	ID129_Run3	325.0943
12	ID128_Run1	394.9231	12	ID138_Run1	306.2549
13	ID129_Run3	393.3585	13	ID128_Run3	297.4167
14	ID126_Run2	377.8679	14	ID126_Run2	296.3922
15	ID128_Run3	374.6078	15	ID62_Run2	288.6154
16	ID62_Run2	349.7115	16	ID128_Run1	284.4286
17	ID62_Run1	328.2245	17	ID62_Run3	277.9792
18	ID62_Run3	327.2917	18	ID62_Run1	266.1633
19	ID18_Run1	314.8980	19	ID18_Run1	260.1837
20	ID123_I10UniXRun2	304.1042	20	ID123_I10UniXRun1	246.9787
21	ID123_I10UniXRun1	295.6250	21	ID123_I10UniXRun2	246.5745
22	ID123_I10UniXRun3	272.5000	22	ID123_I10UniXRun3	232.6744
23	ID124_UNAMiiR12	255.2449	23	ID124_UNAMiiR12	219.1875
24	ID124_UNAMiiR3	139.7021	24	Baseline_mwt	148.2222
25	Baseline_mwt	137.8000	25	ID124_UNAMiiR3	128.3261

Table 6.7: Readability results with the relaxed and strict metric.

is that building an extract of 500 words made of long sentences will be always less informative than a dense coherent summary made of non redundant short sentences. Therefore participants runs had to improve informativeness without hurting readability too much.

Chapter 7

Conclusion

As suggested by Alan Turing, a test to evaluate the ability of a computer to handle a human mind task should involve:

- an interaction with humans where the computer tries to give the illusion that it is human,
- a clear evaluation metric that allows the reproducibility of the experiment,
- a gateway to the open world to explore beyond restricted contexts and closed world assumptions.

Our main motivation relies on the fact that there is no summarization evaluation methodology that encourages research on advanced NLP tasks like summarization by sentence compression. We therefore suggest to come back to Turing's initial motivations: imaging imitation games to answer the controversial philosophical question "do computers have a mind?" without having to define what "mind" means. The question then becomes "what are the common human intellectual tasks that a computer can handle without massive learning?". These are the roots of theoretical computer science where useless tasks for technical applications can be fundamental to understand computers' real limits.

In the original imitation game defined by Turing, there are two players and one assessor. The first player is a human (A) and the second a computer (B). Another human (C) plays the role of the assessor and has to guess the real nature (human or computer) of the two other players. The assessor cannot see the other players, they can just interact with them through a more or less restricted interface that at

least allows to exchange written messages. The assessor asks questions through the interface and has to distinguish the answers given by the human player and those sent by the computer. Turing imagined advanced imitation games to study the spectrum of Artificial Intelligence and compare it to the human mind. However, Turing en-trusted interaction through natural language. In our case, we intend to study the method of interacting itself related to NLP and its linguistic functionalities based on summary generation. Indeed, in the general case of a Turing test, the assessor is not allowed "to see" the players. This is to ensure that they focus on functional aspects and not on appearances. It then seems natural to adapt the imitation game to NLP tasks that try to reproduce human ability to handle texts like summarization and domain mapping. We only considered intellectual tasks that can easily be accomplished by non experts which, in contrast, pose real challenges for an automatic system. We also considered tasks that cannot be carried out without computer assistance like Information Retrieval from large collections.

Bibliography

Agrawal, R., Imielińskivand, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. SIGMOD Rec., 22:207–216.

Amir, A., Aumann, Y., Feldman, R., and Fresko, M. (2005). Maximal Association Rules: A Tool for Mining Associations in Text. Journal of Intelligent Information Systems, 5(3):333–345.

Baeza-Yates, R. and Ribiero-Neto, R. (1999). Modern Information Retrieval. ACM Press, Addison-Wesley.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., and Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In Myaeng, S.-H., Oard, D. W., Sebastiani, F., Chua, T.-S., and Leong, M.-K., editors, SIGIR, pages 667–674. ACM.

Bailey, P., de Vries, A. P., Craswell, N., and Soboroff, I. (2007). Overview of the trec 2007 enterprise track. In Voorhees, E. M. and Buckland, L. P., editors, TREC, volume Special Publication 500-274. National Institute of Standards and Technology (NIST).

Bellot, P., Chappell, T., Doucet, A., Geva, S., Kamps, J., Kazai, G., Koolen, M., Landoni, M., Marx, M., Moriceau, V., Mothe, J., Ramírez, G., Sanderson, M., SanJuan, E., Scholer, F., Tannier, X., Theobald, M., Trappett, M., Trotman, A., and Wang, Q. (2012). Report on inex 2011. SIGIR Forum, 46(1):33–42.

Berry, A., Kaba, B., Nadif, M., SanJuan, E., and Sigayret, A. (2004). Classification et désarticulation de graphes de termes. In Proc. of the 7th International conference on Textual Data Statistical Analysis (JADT 2004), pages 160–170, Louvain-la-Neuve, Belgium.

Berry, A., Pogorelcnik, R., and Simonet, G. (2010). An introduction to clique minimal separator decomposition. Algorithms, 3(2):197–215.

Biha, M., Kaba, B., Meurs, M.-J., and SanJuan, E. (2007). Graph decomposition approaches for terminology graphs. In 6th Mexican International Conference on

Artificial Intelligence (MICAI-07), LNCS 4827, pages 883–893, Aguascalientes, Mexico. Springer.

Braam, R., Moed, H., and A., A. V. R. (1991). Mapping science by combined cocitation and word analysis. 2. dynamical aspects. Journal of the American Society for Information Science, 42(2):252–266.

Cabré, M. (2005). Constituir un corpus de textos de especialidad: condiciones y posibilidades. pages 89–106. Ballard, M.; Pineira-Tresmontant, C. (eds). Arras: Artois Presses Université.

Cabré, M., Bach, C., da Cunha, I., Morales, A., and Vivaldi, J. (2010). Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico. In XXVII Congreso Internacional de AESLA: Modos y formas de la comunicación humana (AESLA 2009). Ciudad Real: Universidad de Castilla-La Mancha.

Cabré, M. (1999). La terminología. Representación y comunicación. Barcelona: IULA-UPF.

Cabré, M. T. (2002). Textos especializados y unidades de conocimiento: metodología y tipologización. In García Palacios, J.; Fuentes, M. T. (eds.). Texto, terminología y traducción. Salamanca: Ediciones Almar, pages 15–36.

Cajolet-Laganière, H. and Maillet, N. (1995). Caractérisation des textes techniques québécois. Présence francophone, (47):113–147.

Callan, J., Croft, W. B., and Harding, S. (1992). The inquery retrieval system. In Proceedings of the 3rd International Conference on Database and Expert Systems Application, pages 78–83.

Callan, J. P., Croft, W. B., and Broglio, J. (1995). Trec and tipster experiments with inquery. Information Processing and Management, 31(3):327 – 343. The Second Text Retrieval Conference (TREC-2).

Callon, T., Courtial, J., and Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Scientometrics, 22(1):155–205.

Castellvi, M. T. C., Bagot, R. E., and Palatresi, J. V. (2001). Automatic term detection: A review of current systems. In Bourigault, D., Jacquemin, C., and L'Homme, M.-C., editors, Recent Advances in Computational Terminology, pages 53–88. John Benjamins.

Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. JASIS, 57(3):359–377.

Chen, C., Ibekwe-Sanjuan, F., and Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. JASIST, 61(7):1386–1409.

Chen, C., Ibekwe-SanJuan, F., SanJuan, E., and Weaver, C. (2006). Visual analysis of conflicting opinions. In 1st International IEEE Symposium on Visual Analytics Science and Technology (VAST 2006), pages 59–66, Baltimore - Maryland, USA.

Chen, H., Wingyan, C., Qin, J., Reid, E., and Sageman, M. (2008). Uncovering the dark web: A case study of jihad on the web. journal of the american society for information science. JASIS, 59(8):1347–1359.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16(1):22–29.

Coulon, R. (1972). French as it is written by French sociologists. Bulletin pédagogique des IUT, (18):11–25.

Cutting, D., Karger, D., Pedersen, J., and Tukey, O. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In 15th Annual International conference of ACM on Research and Development in Information Retrieval - ACM SIGIR, pages 318–329, Copenhagen, Denmark.

da Cunha, I., Cabré, M. T., SanJuan, E., Sierra, G., Moreno, J. M. T., and Vivaldi, J. (2011). Automatic specialized vs. non-specialized sentence differentiation. In Gelbukh, A. F., editor, CICLing (2), volume 6609 of Lecture Notes in Computer Science, pages 266–276. Springer.

da Cunha, I., SanJuan, E., Moreno, J. M. T., Lloberes, M., and Castellón, I. (2012). Diseg 1.0: The first system for spanish discourse segmentation. Expert Syst. Appl., 39(2):1671–1678.

Dang, H. (2008). Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In Proc. of the First Text Analysis Conference.

Denoeud, L., Garreta, H., and Guénoche, A. (2005). Comparison of distance indices between partitions. In et al., P. L., editor, Proceedings of Applied Stochastic Models and Data Analysis, pages 17–20, Brest.

Dobrynin, V., Patterson, D., and Rooney, D. (2004). Contextual document clustering. In Proc. of the 26th European Conference on Information Retrieval (ECIR'04, pages 167–180, Sunderland, UK.

Dunning, T. (1993). Accurate methods for statistics of surprise and coincidence. Computational Linguistics, (19):61–74.

Eguchi, K. and Croft, W. B. (2009). Query structuring and expansion with two-stage term dependence for japanese web retrieval.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. of the National Academy of Science, U.S.A., (95):14863–14868.

Fellbaum, C., editor (1998). WordNet, An Electronic Lexical Database. MIT Press.

Fernández, S., SanJuan, E., and Moreno, J. M. T. (2007). Textual energy of associative memories: Performant applications of enertex algorithm in text summarization and topic segmentation. In MICAI, pages 861–871.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268:2261–2266.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164.

Ganter, B., Stumme, G., and Wille, R., editors (2005). Formal Concept Analysis, Foundations and Applications, volume 3626 of Lecture Notes in Computer Science. Springer.

Geva, S., Kamps, J., and Trotman, A., editors (2010). Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers, volume 6203 of Lecture Notes in Computer Science. Springer.

Glenisson, P., Glänzel, W., Janssens, F., and Moor, B. D. (2005). Combining full text and bibliometric information in mapping scientific disciplines. Information Processing and Management, 41(6):1548–1572.

Gövert, N., Fuhr, N., Lalmas, M., and Kazai, G. (2006). Evaluating the effectiveness of content-oriented xml retrieval methods. Inf. Retr., 9(6):699–722.

Grefenstette, G. (1997). Sqlet: Short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text. In Proceedings of "Recherche d'Information assistée par ordinateur" (RIAO), pages 500–509.

Harris, Z. S. (1968). Mathematical Structures of Language. Wiley, New York.

Hoffmann, L. (1976). Kommunikationsmittel Fachsprache - Eine Einführung. Berlin: Sammlung Akademie Verlag.

Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification, pages 193–218.

Hur, B., Elisseeff, A., and Guyon, I. (2002). A stability-based method for discovering structure in clustered data. Pacific Symposium on Biocomputing, (7):6–17.

Ibekwe, F. and SanJuan, E. (2009). Use of multiword terms and query expansion for interactive information retrieval. In Geva, S., Kamps, J., and Trotman, A., editors, INEX 2008 (selected papers), LNCS 5631, pages 54–64, Berlin Heidelberg. Springer-Verlag.

Ibekwe-SanJuan, F. (1998a). A linguistic and mathematical method for mapping thematic trends from texts. In Proc. of the 13th European Conference on Artificial Intelligence (ECAI), pages 170–174, Brighton, UK.

Ibekwe-SanJuan, F. (1998b). Terminological variation, a means of identifying research topics from texts. In Proc. of Joint ACL-COLING'98, pages 564–570, Québec, Canada.

Ibekwe-SanJuan, F. (2006). Constructing and maintaining knowledge organization tools: a symbolic approach. Journal of Documentation, 62:229–250.

Ibekwe-SanJuan, F. and Dubois, C. (2002). Can syntactic variations highlight semantic links between domain topics? In Proc. of the 6th International Conference on Terminology (TKE), pages 57–63, Nancy, France.

Ibekwe-SanJuan, F. and SanJuan, E. (2003). From term variants to research topics. Journal of Knowledge Organization (ISKO), special issue on Human Language Technology, 29(3/4).

Ibekwe-SanJuan, F. and SanJuan, E. (2004). Mining textual data through term variant clustering: the termwatch system. In Proc. of Recherche d'Information assistée par ordinateur (RIAO), pages 26–28, Avignon, France.

Ibekwe-Sanjuan, F. and SanJuan, E. (2008). Use of multiword terms and query expansion for interactive information retrieval. In Geva, S., Kamps, J., and Trotman, A., editors, INEX, volume 5631 of Lecture Notes in Computer Science, pages 54–64. Springer.

Ibekwe-Sanjuan, F., SanJuan, E., and Vogeley, M. S. E. (2008). Decomposition of terminology graphs for domain knowledge acquisition. In Shanahan, J. G., Amer-Yahia,

S., Manolescu, I., Zhang, Y., Evans, D. A., Kolcz, A., Choi, K.-S., and Chowdhury, A., editors, CIKM, pages 1463–1464. ACM.

Iruskieta, M., de Ilarraza, A. D., and Lersundi, M. (2014). The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In Hajic, J. and Tsujii, J., editors, COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland, pages 466–475. ACL.

Jacquemin, C. (2001). Spotting and discovering terms through Natural Language Processing. MIT Press.

Jain, A. and Moreau, J. (1987). Bootstrap technique in cluster analysis. Pattern Recognition, 20:547–568.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage., 36(6):779–840.

Kageura, K. (2002). The dynamics of Terminology: A descriptive theory of term formation and terminological growth. John Benjamins, Amsterdam.

Kahneman, D., Slovic, P., and Tversky, A. (1981). Judgement under uncertainty - Heuristics and biases. Cambridge University Press, Cambridge.

Kamps, J., Geva, S., Trotman, A., Woodley, A., and Koolen, M. (2008). Overview of the inex 2008 ad hoc track. In PreProceedings of the 15th Text Retrieval Conference (INEX 2008), pages 1–27, Dagstuhl, Germany.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. (2007). Inex 2007 evaluation measures. In Fuhr, N., Kamps, J., Lalmas, M., and Trotman, A., editors, INEX, volume 4862 of Lecture Notes in Computer Science, pages 24–33. Springer.

Karypis, G., Han, E., and Kumar, V. (1994). Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special issue on Data analysis and mining., 32(8):68–75.

Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at jnlpba. In Proc. of JNLPBA-04, pages 70–75.

Knoth, P., Schmidt, M., Smrz, P., and Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods. In ZNALOSTI 2009, Proceedings of the 8th annual conference, page 12. Vydavatelstvo STU.

Kocourek, R. (1991). La langue française de la technique et de la science. Vers une linguistique de la langue savante. Wiesbaden: Oscar Branstetter.

L. Denoyer, P. G. (2006). The wikipedia xml corpus. In SIGIR Forum, page 6.

Lalmas, M. and Tombros, A. (2007). Evaluating xml retrieval effectiveness at inex. SIGIR Forum, 41(1):40–57.

L'Homme, M. (1993). Contribution à l'analyse grammaticale de la langue d'espécialité : le mode, le temps et la personne du verbe dans quelques textes, scientifiques écrits à vocation pédagogique. Québec: Université Laval.

L'Homme, M. (1995). Formes verbales de temps et texte scientifique. Le langage et l'homme, 2-3(31):107–123.

Lin, Z., Chua, T.-S., Kan, M.-Y., Lee, W. S., Qiu, L., and Ye, S. (2007). NUS at DUC 2007: Using Evolutionary Models of Text.

Louis, A. and Nenkova, A. (2009). Performance confidence estimation for automatic summarization. In EACL, pages 541–548. The Association for Computer Linguistics.

Mani, I. and Mayburi, M. (1999). Advances in automatic text summarization. The MIT Press, U.S.A.

Marchionini, G. (1992). Interfaces for end-user information seeking. JASIS, 43(2):156–163.

McKeown, K. and Radev, D. (1995). Generating summaries of multiple news articles. In 18^{th} ACM SIGIR, pages 74–82.

Metzler, D. and Croft, W. B. (2003). Combining the language model and inference network approaches to retrieval. Information Processing and Management, 40(5):735–750.

Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. Inf. Process. Manage., 40(5):735–750.

Metzler, D. and Croft, W. B. (2005). A markov random field model for term dependencies. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 472–479.

Metzler, D. and Croft, W. B. (2007). Latent concept expansion using markov random fields. In Proceedings of the 30th Annual International ACM SIGIR Conference, pages 311–318, New York, NY. ACM, ACM.

Metzler, D., Lavrenko, V., and Croft, W. B. (2004). Formal multiple-bernoulli models for language modeling. In Proceedings of SIGIR '04, number 540-541. a poster presentation.

Metzler, D., Strohman, T., Turtle, H., and Croft, W. B. (2005). Indri at trec 2004: Terabyte track. page electronic proceedings only.

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, page 20, Morristown, NJ, USA. Association for Computational Linguistics.

Miller, G. A. (1994). Wordnet: A Lexical Database for English. In HLT. Morgan Kaufmann.

Milligan, G. W. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. Pychometrika, 50:159–179.

Milligan, G. W. and Cooper, M. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioural Research, 21:441–458.

Mishne, G. and de Rijke, M. (2006). Boosting web retrieval through query operations. Lecture Notes in Computer Sciences, 3408:502 – 516.

Moriceau, V., SanJuan, E., Tannier, X., and Bellot, P. (2009). Overview of the 2009 qa track: Towards a common task for qa, focused ir and automatic summarization systems. In Geva et al. (2010), pages 355–365.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In Proceedings of HLT-NAACL, volume 2004.

Ng, R. and Han, J. (2002). Clarans: A method for clustering objects or spatial data mining. In IEEE transactions on knowledge and data engineering, volume 14.

Pantel, P. and Lin, D. (2002). Clustering by committee. In Annual International conference of ACM on Research and Development in Information retrieval - ACM SIGIR, pages 199–206, Tampere, Finland.

Perez-Carballo, J. and Strzalkowski, T. (2000). Natural language information retrieval: progress report. Information Processing and Management, 36(1):155 – 178.

Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In ACL, pages 544–554.

Polanco, X., Grivel, L., and Royauté, J. (1995). How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators. In Proc. of the 5th International Conference of the International Society for Scientometrics and Informetrics, pages 435–444, Illinois, U.S.A.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 275–281, New York, NY, USA. ACM.

Porter, M. (2006). An algorithm for suffix stripping. Program: electronic library and information systems, 40(3):211–218.

Price, L. and Thelwall, M. (2005). The clustering power of low frequency words in academic webs. Journal of the American Society for Information Science and Technology, 56(8):883–888.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In SIGIR, pages 213–220. ACM.

Saggion, H., Moreno, J. M. T., da Cunha, I., SanJuan, E., and Velázquez-Morales, P. (2010). Multilingual summarization evaluation without human models. In Huang, C.-R. and Jurafsky, D., editors, COLING (Posters), pages 1059–1067. Chinese Information Processing Society of China.

SanJuan, E. (2011). Mapping knowledge domains - combining symbolic relations with graph theory. In Filipe, J. and Fred, A. L. N., editors, KDIR, pages 527–536. SciTePress.

SanJuan, E., Bellot, P., Moriceau, V., and Tannier, X. (2010). Overview of the inex 2010 question answering track (qa@inex). In Geva, S., Kamps, J., Schenkel, R., and Trotman, A., editors, INEX, volume 6932 of Lecture Notes in Computer Science, pages 269–281. Springer.

Sanjuan, E., Dowdall, J., Ibekwe-Sanjuan, F., and Rinaldi, F. (2005). A symbolic approach to automatic multiword term structering. Computer Speech Language (CSL), 19(4):524–542.

SanJuan, E., Flavier, N., Bellot, P., and Ibekwe-Sanjuan, F. (2008). Universities of avignon and lyon iii at trec 2008: Enterprise track. In Voorhees, E. M. and Buckland, L. P., editors, TREC, volume Special Publication 500-277. National Institute of Standards and Technology (NIST).

SanJuan, E. and Ibekwe-Sanjuan, F. (2006). Text mining without document context. Inf. Process. Manage., 42(6):1532–1552.

SanJuan, E. and Ibekwe-Sanjuan, F. (2009). Combining language models with nlp and interactive query expansion. In Geva et al. (2010), pages 122–132.

Schenkel, R., Suchanek, F. M., and Kasneci, G. (2007). Yawn: A semantically annotated wikipedia xml corpus. In Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., and Brochhaus, C., editors, BTW, volume 103 of LNI, pages 277–291. GI.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing, volume 12. Manchester, UK.

Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, (19):143–177.

Smeaton, A. F. (1999). Using nlp and nlp resources for information retrieval tasks. pages 99–109. Kluwer Academic Publishers.

Sparck-Jones, K. (1999). What is the role for nlp in text retrieval. In Strzalkowski, T., editor, Natural language information retrieval, pages 1–25. Kluwer Academic Publishers.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. (2005). Indri: A language-model based search engine for complex queries (extended version). IR 407, University of Massachusetts.

Strzalkowski, T., Carballo, J. P., Karlgren, J., Hulth, A., Tapanainen, P., and Lahtinen, T. (1999). Natural language information retrieval: Trec-8 report. In TREC.

Tibshirani, R., Walther, G., and Hastie, T. (2000). Estimating the number of clusters in a dataset via the gap statistic. In Technical Report, number 208, Dept. of Statistics, Stanford University.

Tversky, A. and Kahneman, D. (1990). Judgment under uncertainty: heuristics and biases. pages 32–39.

Vechtomova, O. (2005). The role of multi-word units in interactive information retrieval. In Losada, D. E. and Fernández-Luna, J. M., editors, ECIR, volume 3408 of Lecture Notes in Computer Science, pages 403–420. Springer.

Vivaldi, J. (2009). Corpus and exploitation tool: IULACT and bwanaNet. In I International Conference on Corpus Linguistics (CICL-09), pages 224–239. Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) A survey on corpus-based research, Universidad de Murcia.

Voorhees, E. M. (1999). Natural language processing and information retrieval. Lecture Notes in Computer Sciences, 1714:32 – 48.

Watcholder, N., Evans, D., and Klavans, J. (2001). Automatic identification of index terms for interactive browsing. In Proceedings of the ACM IEEE Joint Conference on Digital libraries, pages 116 – 124, Roanoke, Virginia.

Weeds, J., Dowdall, J., an B. Keller, G. S., and Weir, D. (2005). Using distributional similarity to organise biomedical terminology. Terminology: Special Issue on Application-driven terminology engineering, 11(1):107–141.

Wehrens, R., Buydens, L. M., Fraley, C., and Raftery, A. E. (2003). Model-based clustering for image segmentation and large datasets via sampling. Technical Report 424, Department of Statistics, University of Washington.

Yeung, K. and Ruzzo, W. (2001). Details of the adjusted rand index and clustering algorithms. supplement to the paper "an experimental study on principal component analysis for clustering gene expression data. Bioinformatics, (17):763–774.

Zaki, M. J. (2009). Closed itemset mining and non-redundant association rule mining. In Liu, L. and Özsu, M. T., editors, Encyclopedia of Database Systems, pages 365–368. Springer US.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst., 22(2):179–214.

Zitt, M. and Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. Scientometrics, 30(1):333–351.