Language Independent Methodologies to Tackle Multilinguality

Gaël Dias, Célia Nunes, João Paulo Cordeiro, Rumen Moraliyski, Isabel Marcelino, Raycho Mukelov, Ricardo Campos, Cláudia Santos, Elsa Alves, Bruno Conde and Bono Nonchev

Centre for Human Language Technology and Bioinformatics, Department of Computer Science, University of Beira Interior, Covilhã, Portugal
{ddg, celia, jpaulo, rumen, isabel, raicho, ricardo, claudia, elsalves, bruno, bono}@hultig.di.ubi.pt

Abstract

Until now, Natural Language Processing (NLP) research development has mainly been conducted for the English speaking community. However, the European Union with its 25 member-states already involves 22 different official languages. As a consequence, multilinguality is certainly the most important challenge of this century for the European NLP community. In this paper, we show how the Centre for Human Language Technology and Bioinformatics has been dealing with the problem of multilinguality by proposing language independent systems instead of language tailored architectures.

1. Introduction

In the beginning of this century, multilingual projects appear on the schedule as the European Union with its 25 member-states is now a reality and is facing major social, economic and political challenges in order to become a world wide driving force. In fact, language diversity acts as a barrier preventing absolute free trading within Europe. But, if this challenge can be overcome, this apparent weakness will prove to be Europe's strongest proof of unity. So, it is clear that multilinguality is certainly the most important challenge of this century for the European community in the field of Natural Language Processing. Otherwise, we may observe a two speed European Union: rich countries and the rest. However, until now, research development has mainly been conducted for the English speaking community, and a number of a priori methodologies have emerged. First, it was Chomsky's theoretical background that was applied to Natural Language Processing, with known unsatisfactory results. Then, a sharp shift of methodologies occurred. Some Statistics came into the area, but used methodologies required huge quantities of manually annotated corpora. But, no one ever questioned the convenience of such annotated corpora, marked-up according to the knowledge we thought we had about human languages. As a result research evolved as well as the methods used. However, only a few researchers tried to push the ever growing application of statistical methods to the extreme where any text, widely available in the web, could be used as it is, without any annotation, to automatically learn from it. With such an approach, we may overcome problems that, all over Europe, researchers from smaller countries, with smaller populations, could feel, mainly due to the lack of available linguistic resources such as annotated corpora, shallow parsers and so on and so forth. This other perspective where statistics inference must be heavily used in order to enable computers to learn from the web huge amount of texts, with as little language knowledge as possible, will allow a faster breaking of human languages code while producing a rigorous scientific treatment of data without any a priori imposition of human "knowledge" or "ignorance" about human languages. That is where the Centre for Human Language Technology and Bioinformatics wants to intervene on the European scene gathering these few researchers who believe in a new way of treating human languages. In this paper, we will present many systems that use as less as possible linguistic resources so that they can be applied to a great deal of languages and benefit as much as possible multilingual systems and projects.

2. Multiword Lexical Unit Extraction

The acquisition of terminologically relevant multiword lexical units from large text collections is a fundamental issue in the context of Information Retrieval. Indeed, their identification leads to improvements in the indexing process and allows guiding the user in his search for information. On one hand, selecting discriminating terms in order to represent the contents of texts is a critical problem. Ideally, the indexing terms should directly describe the concepts present in the documents. However, most of the information retrieval systems index the documents of a text collection based on individual words that are not specific enough to evidence the contents of texts. In order to improve the quality of the indexing process, some systems take advantage of pre-existing thesauri. In that case, the discriminating terms are selected from the thesaurus (Betts and Marrable, 1991). Unfortunately, most of the domains do not contain pre-defined thesauri and very few projects include automatic construction of specialised thesauri (Grefenstette, 1994). In order to overcome the lack of domain specific thesauri, evolutionary retrieval systems use multiword terms previously extracted from text collections to represent the contents of texts (Evans and Lefferts, 1993). Indeed, multiword terms embody meaningful sequences of words that are less ambiguous than single words and allow approximating more accurately the contents of texts. However, most of the multiword terms are not listed in lexical databases. Indeed, the creation, the maintenance and the upgrade of terminological data banks often require a great deal of manual efforts that cannot cope with the ever growing number of text corpora to analyse. Moreover, due to the constant dynamism of specialised languages, the set of multiword terms is opened and to be completed. Indeed, most of the neologisms in technical and scientific domains are realised by multiword terms. For example, *World Wide Web*, *IP address* and *TCP/IP network* are terminologically relevant multiword lexical units that are particularly new in the domain of Computer Science. As a consequence, there has been a growing interest in developing techniques for automatic term extraction. In order to extract multiword terms from text corpora, three main strategies have been proposed in the literature. First, purely linguistic systems (David and Plante, 1990; Bourigault, 1996) propose to extract relevant terms by using techniques that analyse specific syntactic structures in the texts. However, this methodology suffers form its monolingual basis, as the systems require highly specialised linguistic techniques to identify clues that isolate possible candidate terms. Second, hybrid methodologies (Justeson, 1993; Daille, 1995) define co-occurrences of interest in terms of syntactic patterns and statistical regularities. However, by reducing the searching space to groups of words that correspond to *a priori* defined syntactic patterns (Noun+Adj, Noun+Prep+Noun etc...), such systems do not deal with a great proportion of terms and introduce noise in the retrieval process. Finally, purely statistical systems (Church and Hanks, 1990; Dunning, 1993; Smadja, 1993) extract discriminating multiword terms from text corpora by means of association measure regularities. As they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and extract relevant units independently from the domain and the language of the input text. However, they emphasise two major drawbacks. On one hand, by relying on *ad hoc* establishment of global thresholds they are prone to error. On the other hand, as they only allow the acquisition of binary associations, these systems must apply enticement techniques to acquire multiword terms with more than two words. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams for the initiation of the iterative process. In order to overcome the problems previously highlighted by the statistical systems, we propose a new architecture called SENTA (Software for the Extraction of N-ary Textual Associations) which conjugates a new association measure called the Mutual Expectation (Dias, 2002) with a new acquisition process called the GenLocalMaxs (Dias, 2002). On one hand, the Mutual Expectation, based on the concept of Normalised Expectation, evaluates the degree of cohesiveness that links together all the textual units contained in an n-gram (i.e. $\forall n, n \geq 2$). On the other hand, the GenLocalMaxs retrieves the candidate terms from the set of all the valued n-grams by evidencing local maxima of association measure values. The combination of the new association measure with the new acquisition process proposes an innovative integrated solution to the problems of enticement techniques and global thresholds defined by experimentation. This system can be freely downloaded at http://senta.di.ubi.pt. In particular, we show in the next section that it allows improved result for unsupervised topic segmentation.

3. Unsupervised Topic Segmentation

Topic segmentation is the task of breaking documents into topically coherent multi-paragraph subparts. In particular, topic segmentation has extensively been used in text summarization where it serves as the basic text structure in order to apply sentence extraction and sentence compression techniques (Angheluta et al., 2002). However, most methodologies are based on lexical repetition that show evident reliability problems or rely on harvesting linguistic resources which are usually available for dominating languages. In order to tackle these drawbacks, we developed an innovative topic segmentation system based on a new informative similarity measure that takes into account word co-occurrence in order to avoid the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases, and evaluate it on a set of web documents belonging to a single domain. In particular, our architecture solves three main problems evidenced by previous research. First, systems based uniquely on lexical repetition show reliability problems (Hearst, 1994; Reynar, 1994; Sardinha, 2002) as common writing rules prevent from using lexical repetition. Second, systems based on lexical cohesion, using existing linguistic resources that are usually only available for dominating languages like English, French or German do not apply to less favoured and emerging languages (Morris and Hirst, 1991; Kozima, 1993). Third, systems that need previously existing harvesting training data (Beeferman et al., 1997) do not adapt easily to new domains as training data is usually difficult to find or build depending on the domain being tackled. Instead, our architecture proposes a language-independent unsupervised solution, similar to (Phillips, 1985; Ponte and Croft, 1997), defending that topic segmentation should be done "on the fly" on any text thus avoiding the problems of domain, genre, or language-dependent systems. Our algorithm is based on the vector space model which determines the similarity of neighbouring groups of sentences and places subtopic boundaries between dissimilar blocks. In our specific case, each sentence in the corpus is evaluated in terms of similarity with the previous block of *k* sentences and the next block of *k* sentences. According to us, two main factors must be taken into account to define the relevance of a word for the specific task of Topic Segmentation as shown in (Dias and Alves, 2005): its semantic importance and its distribution across the text. Once each word has been evaluated, the next step of the application of the vector space model aims at determining the similarity of neighbouring groups of sentences. For that purpose, we propose a new informative similarity measure, the *infosimba* measure that includes in its definition the Equivalence Index Association Measure (*EI*) proposed by (Muller et al., 1997), so that word co-occurrence information is included in the evaluation of similarity. Finally, placing subtopic boundaries between dissimilar blocks is performed based on the standard deviation algorithm proposed by (Hearst, 1994) and the definition of a score for each sentence exactly as (Beeferman *et al.*, 1997) compare short and long-range models. In order to be as complete as possible, we ran the c99 algorithm (Choi, 2000), the TextTiling algorithm (Hearst, 1994) and our algorithm on a benchmark that gathers texts from three different family languages (English, Portuguese and Bulgarian) and from which multiword units have been identified. The first astonishing result is that the c99 algorithm is the one that performs the worst over our test corpus. This goes against Chois's (2000) evaluation that evidences improved results when compared to the TextTiling algorithm over the c99 corpus. This result clearly shows that the c99 cannot be taken as a gold standard for topic segmentation evaluation schemes. The reason why the TextTiling algorithm performs better than the c99 on our benchmark is the fact that (Hearst, 1994) uses the appearance of new lexical units as a clue for topic boundary detection whereas (Choi, 2000) relies more deeply on lexical repetition which is drastically penalized. The second result has to do with the evaluation metrics. In particular, the P_k estimate (Beeferman et al., 1997) gives better results for the c99 than for the TextTiling although the F-measure and the *Windowdiff* (Pezner and Hearst, 2002) show the contrary. This result confirms the conclusions of (Pevzner and Hearst, 2002) about the fact that the P_k "penalizes false negatives more heavily than false positives, over-penalizes near misses and is affected by variation in segment size distribution". However, the *WindowDiff* also shows experimental problems. In particular, for the case with Multiword Unit identification, while the F-measure and the P_k estimate clearly show better results for our system than for the TextTiling algorithm, the *WindowDiff* shows opposite results. The problem evidenced here is the fact that the *WindowDiff* over-evaluates near misses. So, none of the three evaluation metrics show reliable results, as the F-measure also does not differentiate near misses from far misses. As the introduction of Multiword Units is concerned, only the c99 algorithm seems to be insensitive to this phenomenon. Indeed, while our algorithm and the TextTiling greatly benefit of the identification of Multiword Units (6% improvement for the F-measure for the TextTiling algorithm and 5% improvement for our algorithm), the c99 shows unchanged results. Finally, our system shows better results than both systems in the case where Multiword Units are previously extracted. In particular, it shows 30% improvement over the c99 algorithm and 21% over the TextTiling with respect to the F-measure. The same conclusion can be drawn when Multiword Units are not extracted. In this case, our algorithm shows 25% improvement over the c99 algorithm and 22% over the TextTiling with respect to the F-measure. In order to be complete, we will talk about some astonishing results that occurred with Bulgarian. In fact, for Bulgarian, our algorithm looses 7% of F-measure when multiword units are introduced in the texts. However, this is not the case for TextTiling that, nevertheless, does not show any improvement when compared to single word evaluation. These figures were quite surprising at first sight. However, after some deeper analysis of the results, we came to the conclusion that SENTA (Dias, 2002) was identifying too many locutions and not enough compounds. In fact, due to the morphology of Bulgarian that accepts many derivations for one and the same concept word (i.e. a word with strong semantic value), SENTA elects more locutions that usually show syntactic phenomena than concepts like compound nouns or verbs that are semantically strong. As a consequence, Multiword locutions are over-evaluated in the process of topic segmentation and wrongly induce the topic boundary detection. The system and its evolutions will soon be available at the following address: http://asas.di.ubi.pt. In the next section, we propose to extract lexical chains from part-of-speech texts. This comes as a natural follow up of topic segmentation systems as shown in (Barzilay and Elhadad, 1997).

4. Extraction of Lexical Chains

Lexical chains are powerful representations of documents compared to broadly used bag-of-words representations. In particular, they have successfully been used in the field of automatic text summarization (Barzilay and Elhadad, 1997). However, until now, lexical chaining algorithms have only been proposed for English as they rely on linguistic resources such as Thesauri (Morris and Hirst, 1991) or Ontologies (Barzilay and Elhadad, 1997; Silber and McCoy, 2002; Galley and McKceown, 2003). (Morris and Hirst, 1991) were the first to propose the concept of Lexical Chains to explore the discourse structure of a text. However, at the time of writing their paper, no machine-readable thesaurus was available so they manually generated lexical chains using Roget's Thesaurus. A first computational model of Lexical Chains is introduced by (Hirst and St-Onge, 1997). Their biggest contribution to the study of Lexical Chains is the mapping of WordNet relations and paths (transitive relationships) to (Morris and Hirst, 1991) word relationship types. However, their greedy algorithm does not use a part-of-speech tagger. Instead, the algorithm only selects those words that contain noun entries in WordNet to compute lexical chains. But, as (Barzilay and Elhadad, 1997) point out, the use of a part-of-speech tagger could eliminate wrong inclusions of words such as “read”, which has both noun and verb entries in WordNet. So, they propose the first dynamic method to compute lexical chains. They argue that the most appropriate sense of a word can only be chosen after examining all possible lexical chain combinations that can be generated from a text. Because all possible senses of the word are not taken into account, except at the time of insertion, potentially pertinent context information that is likely to appear after the word is lost. However, this method of retaining all possible interpretations until the end of the process causes the exponential growth of the time and space complexity. As a consequence, (Silber and McCoy, 2002) propose a linear time version of (Barzilay and Elhadad, 1997) lexical chaining algorithm. In particular, their implementation creates a structure, called meta-chains, that implicitly stores all chain interpretations without actually creating them, thus keeping both the space and time usage of the program linear. Finally, (Galley and McKeown, 2003) propose a chaining method that disambiguates nouns prior to the processing of lexical chains. Their evaluation shows that their algorithm is more accurate than (Barzilay and Elhadad, 1997; Silber and McCoy, 2002) ones. One common point of all these works is that lexical chains are built using WordNet as the standard linguistic resource. Unfortunately, systems based on static linguistic knowledge bases are limited. First, such resources are difficult to find. Second, they are largely obsolete by the time they are available. Third, linguistic resources capture a particular form of lexical knowledge which is often very different from the sort needed to specifically relate words or sentences. In particular, WordNet is missing a lot of explicit links between intuitively related words. (Fellbaum, 1998) refers to such obvious omissions in WordNet as the ”tennis problem” where nouns such as “nets”, “rackets” and “umpires” are all present, but WordNet provides no links between these related tennis concepts. In order to solve these problems, we propose to automatically construct from a collection of documents a lexico-semantic knowledge base with the purpose to identify cohesive lexical relationships between words based on corpus evidence (Dias et al., 2006). This hierarchical lexico-semantic knowledge base is built by using the Pole-Based Overlapping Clustering Algorithm (Cleuziou et al., 2003) that clusters words with similar meanings and allows words with multiple meanings to belong to different clusters. The second step of the process aims at automatically extracting Lexical Chains from texts based on our knowledge base. For that purpose, we propose a new greedy algorithm which can be seen as an extension of (Hirst and St-Onge, 1997) and (Barzilay and Elhadad, 1997) algorithms. In particular, it implements (Lin, 1998) information-theoretic definition of similarity as the relatedness criterion for the attribution of words to lexical chains. Our experimental evaluation shows that relevant lexical chains can be constructed with our lexical chaining algorithm. Indeed, comparatively to (Barzilay and Elhadad, 1997) algorithm, we produce longer and more meaningful lexical chains and much less occurrences of lexical chains with only one word – this characteristic is evident for (Barzilay and Elhadad, 1997) algorithm where a great deal of lexical chains only contain one word. However, we acknowledge that more comparative evaluations must be done in order to draw definitive conclusions. The system will soon be available at http://alexia.di.ubi.pt. Although the soft clustering algorithm proposes an interesting starting point to build a knowledge-base, some work still has to be done to produce quality ontology. For that purpose, we have recently undergone work in the context of synonym detection from corpora.

5. Detection of Synonyms

Many repetitions of a word in the same text are unpleasant for the reader. In order to ease the reading, synonymy is normally used to refer the same concept within short distance. Sometimes metonymic collocations are involved in the role of synonymy. This richness and creativity of language poses many problems when dealing exclusively with statistical lexical analysis. For example, it makes difficult to calculate similarity measure between texts (sentence, paragraph, text) due to the lack of lexical repetition. To deal with those obstacles, thesauri and resources like WordNet have been developed. Since language is dynamic, systems which provide semantic data are out-of-date when they are needed and as a consequence need constant updating. Even when up-to-date resources are available, they are usually general, and particular to a domain and suffer low coverage. This is why we aim at developing a method for automatic discovery of synonymy relations between words. For a reference and starting point, we chose a set of TOEFL test cases, used in (Turney et al., 2003). Those cases comprise 5 words each – the target one, the correct answer i.e. the synonym and 3 more less relevant words. Current stages of our study deal only with the noun cases from this test set. Our method is based on the presupposition that reluctance towards repetition applies both at phrase and/or at lexical level. According to Harris distributional hypothesis two similar words are expected to have similar distributions over their contexts. Since noun-verb and verb-noun distributional constructions convey most of the information of a sentence, the same verb coupled with synonyms of a same noun would express the same idea. Hence, when considered within a single text similar distributions of synonyms over their verbs would mean repetition of ideas. Thus, we expect that even though two synonyms have similar distributions over their contexts throughout a large amount of text, when considered within the limits of one document they are expected to be near perpendicular, said in terms of vector space model. Thus, our model relies on the global similarity and the intra-textual dissimilarity of word distributions to estimate their degree of synonymy. In order to substantiate this idea we needed corpora from which to gather statistics for word pairs and their contexts provided that both words appear in the same text. We queried Google™ with 4 pairs of words for each TOEFL test case – the target word together with one of the other 4 in the case – and collected all of the afforded web pages, striped out the HTML tags and lemmatized/shallow parsed the remaining texts. For the linguistic treatment, we used the MontyLingua library developed by (Liu, 2004). From the whole corpora gathered we selected those documents that contained 3 or more times both words from at least one of the test pairs. Thus, the corpus consists of 38.794.161 words and 122.665 distinct word tokens. From these automatically built resources, we will test our hypothesis using different methods for word distributional representation as mentioned in (Baroni and Brisi, 2004) with different text information (raw text, part-of-speech tagged text or shallow parsed text). This work aims at structuring our knowledge base so that we can produce better text summarization systems based on lexical chains and topic segmentation. Finally, in the next section, we present a new text summarization paradigm.

6. Sentence Compression

A recent and relevant summarization topic is sentence compression, which aims to go a step further, beyond simple sentence extractive summarization techniques. As the title suggests, the focus is targeted on the sentence, rather than the whole text. Even at a sentence level, we may have some long and complex structures, with a considerable amount of "superfluous" components like prepositional phrases, adjectival or adverbial elements. In most cases, the volume of new information added by these "spurious" items is negligible. For that purpose, different techniques for sentence simplification have been proposed, either by content cutting or content transformation. A set of approaches have been experimented in this field, some using statistics and machine learning tools, others using huge language knowledge elements like thesauri. For instance, with machine learning techniques, two supervised methods were tried in (Knight and Marcu, 2002) – the noisy channel model and the decision-based model. A practical difficulty inherent to the supervised algorithms consists in the need for supplying training examples, which are costly in this domain. For example, in the work referred previously, 1057 pairs of sentences were used for training. Considering the size of a language in terms of its sentences and all possible combinations, one may criticize that such a data set is very small and unable to model the whole variety of possible sentence transformations. Approaches using language knowledge resources were also experimented in sentence compression. An example is detailed in (Jing and Mckeown, 2000) where important language resources were employed, like syntactic dictionaries, English verb classes, alternations and even the Brown corpus tagged with WordNet senses. Although such strategies may achieve good results, they are strongly dependent upon the existence of language resources, which are abundant for English and other major languages but scarce or even inexistent for many others. Considering the difficulties referred previously, there exists space and need for a new kind of research in this field, tackling the language independency and overcoming the training difficulties inherent to supervised learning algorithms. Therefore, we propose a new approach that consists in the following main steps: (a) extraction of pairs of non-symmetric entailed paraphrases, from corpora (for example: web news stories, very abundant nowadays), (b) automatic alignment of the extracted paraphrase to construct a huge training dataset and (c) induction of sentence compression rules, by applying learning mechanisms. For the first step, we proposed a new metric called the Sumo-metric (Cordeiro et al., 2007) to automatically find paraphrases in text, and compared it with the most widely used in this domain: the edit-distance, the word n-gram overlap, and the BLEU metric (Papineni et al., 2001). Experimentation showed that the Sumo-metric achieves better results than any other tested metric. We concluded that our metric is well tailored for paraphrase detection in corpora, especially to avoid pairs that are almost equal sentences. Those pairs are obviously useless for sentence compression. We are also engaged with the automatic construction of a huge paraphrase corpus that may be used in a wide variety of research fields, like automatic text generation, for instance. So far, the only one available is the Microsoft Research Paraphrase Corpus (Dolan et al., 2004), with 3900 paraphrases, extracted from news stories and selected by humans. In step (b) there exists a set of alignment algorithms widely tested on the field of automatic text translation that may be adapted for paraphrase alignment. However, we are also interested in applying the algorithm developed by (Doucet and Ahonen-Myka, 2006) to align clusters of paraphrases. Finally, step (c) is under research and we will consider the best options available, from machine learning, for compression rule induction. Unlike (Knight and Marcu, 2002) who propose sub-symbolic knowledge like statistical models, we aim at producing symbolic knowledge using the Inductive Logic Programming paradigm over texts. However, in this case, we will need at least part-of-speech tagged corpora. More information can be found at http://competence.di.ubi.pt.

7. Web Search Results Clustering

While the first part of this paper was devoted to text summarization and construction of ontology, we also investigated web search systems. In particular, current search engines return lists of ranked urls with their title and a short description of the document, known as snippets. However, users still deal with the problem of finding relevant web pages between the lists of retrieved results. One of the main problems is that the induced relevance defined by the search engines may not satisfy the user's needs. Although search engines are useful, they fail to present the results in an appropriate manner, thus making difficult to the user to find the appropriate information he is looking for during the browsing process. Based on these observations, some new commercial search engines have appeared in the last years. Some of the most relevant examples are *Vivissimo*, *iBoogie*, *Clusty* and *Grokker*. Each one of these search engines builds a set of labeled hierarchical clusters processed on the fly over web snippets, a process also known as post-retrieval document browsing or ephemeral clustering. This process can be seen as a bottom-up process, with the categories being part of the output, rather than part of the input (Maarek et al., 2000). Indeed, in this case, clusters do not require pre-defined categories, such as in classification methods (Zeng et al., 2004). Hierarchical clustering of web pages is today a relevant problem in Information Retrieval (IR). As (Ferragina and Gulli, 2003) claim, it is an innovative approach to help users searching for relevant web pages, otherwise undiscovered because of their location in the ranking, and seems to be the PageRank of the future. In this scope, and aside commercial solutions, where none or little information is available, some scientific literature has been published, but all have ignored the potential of using web content mining techniques to semantically analyze a web page. Without this analysis, systems are not prepared to understand completely the contents of documents. As a consequence, the ambiguity problem (query term may have more than one meaning) and the synonymy problem (web documents may only have just synonymy of the query term) remain unsolved. This issue is a central problem in the context of modern IR and tends to get worse when the user is not familiar with the topic he is searching for. To tackle these drawbacks, we developed a meta-search engine called WISE (Campos and Dias, 2005). Through the use of web content mining techniques introduced in the context of the Webspy software (Veiga et al., 2004) and statistical methodologies for phrase detection with the SENTA software (Dias, 2002) to semantically represent the content of web documents, the system, which is a web-based interface generates soft hierarchical clusters on the fly, without pre-defined groups or pre-built knowledge bases, by applying an overlapping clustering algorithm called PoBOC (Cleuziou et al., 2003). In particular, the PoBOC algorithm, which is graph based, allows a document to be in multiple clusters (overlap), reflecting the fact that a web page may contain different meanings of the query terms. We believe that our solution is innovative as the architecture as a whole, and not just part of it, is language and topic independent and as a consequence is real-world web adaptable unlike most of the methodologies proposed so far. As a whole, WISE is a web search interface system, allowing the user to choose which search engine will run the query. In response to the query, the system returns a page with a set of clusters and their associated key concepts which are keywords representing the web documents. Below each keyword there exists a list of urls, in one or more clusters, so that the user can easily choose the web page he wants to see. Our algorithm is composed of five steps: (1) Search results gathering; (2) Selection of relevant web pages; (3) Document parsing for phrase extraction; (4) Document parsing for key concept extraction; (5) Hierarchical clustering and labeling. As a result, we propose a structured - indexed catalogue of retrieved urls instead of an ordered list of relevant documents. Experimental results demonstrate correctness of the clusters, the appropriate quality and descriptiveness of the labels, concept disambiguation and language-independence. In particular, the system will soon be available on http://wise.di.ubi.pt.

8. Conclusions and Future Work

In this paper, we have shown many different research directions to deal with multilingual systems by providing language independent architectures. It is clear that some linguistic resources can/must be added to these methodologies, but in all cases they should also be acquired automatically to prevent any a priori imposition of human "knowledge" or "ignorance" about human languages.

References

(Angheluta et al., 2002) Angheluta, R., De Busser, R., Moens, M-F. 2002. The Use of Topic Segmentation for Automatic Summarization. In Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization. July 11-12, Philadelphia, Pennsylvania, USA.

(Baroni and Bisi 2004) Baroni, M. and Bisi, S. 2004. Using coocurence statistics and web to discover synonyms in a technical language. In Proceedings of LREC 2004, Lisbon: ELDA.1725-1728.

(Barzilay and Elhadad, 1997) Barzilay R. and Elhadad M. 1997. Using Lexical Chains for Text Summarization. Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS-97), ACL, Madrid, Spain.10--18.

(Beeferman et al., 1997) Beeferman, D., Berger, A., and Lafferty, J. 1997. Text segmentation using exponential models. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, 35--46.

(Betts and Marrable, 1991) Betts, R. and Marrable D. 1991. Free Text vs controlled vocabulary, retrieval precision and recall over large databases. In Online Inf 91. London. 153--165.

(Bourigault, 1996) Bourigault, D. 1996. Lexter, a Natural Language Processing Tool for Terminology Extraction. In Proceedings of 7th EURALEX International Congress.

(Campos and Dias, 2005) Campos, R. and Dias, G. 2005. Automatic Hierarchical Clustering of Web Pages. In Proceedings of the ELECTRA Workshop associated to the 28th Annual International ACM SIGIR Conference, Salvador, Brazil, August 19, 83--85.

(Choi, 2000) Choi, F.Y.Y. 2000. Advances in Domain Independent Linear Text Segmentation. In Proceedings of NAACL'00, Seattle, April 2000. ACL.

(Church and Hanks 1990) Church, K.W. and Hanks P. 1990. Word Association Norms Mutual Information and Lexicography. In Computational Linguistics, 16 (1). 23--29.

(Cleuziou et al., 2003) Cleuziou, G., Martin, L. and Vrain, C. 2003. PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In Proceedings of the 16th ECAI, Valencia, Spain, 440--444.

(Cordeiro et al., 2007) Cordeiro, J.P., Dias, G. and Brazdil, P. 2007. Unsupervised Learning of Paraphrases. In Research in Computer Science. National Polytechnic Institute, Mexico. ISSN 1870-4069. To appear.

(David and Plante, 1990) David, S. and Plante, P. 1990. Termino Version 1.0. Research Report of Centre d'Analyse de Textes par Ordinateur. Université du Québec. Montréal.

(Daille, 1995) Daille, B. 1995. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In The balancing act combining symbolic and statistical approaches to language. MIT Press.

(Dias, 2002) Dias, G. 2002. Extraction Automatique d'Associations Lexicales à Partir de Corpora. PHD Thesis, New University of Lisbon (Portugal) and University of Orléans

(France).

(Dias and Alves, 2005) Dias, G. and Alves, E. 2005. Unsupervised Topic Segmentation Based on Word Co-occurrence and Multi-Word Units for Text Summarization. In Proceedings of the ELECTRA Workshop associated to 28th Annual International ACM SIGIR Conference, Salvador, Brazil, August 19. pp. 41-48. In association with ACM editions. ISBN: 1595930345.

(Dias et al., 2006) Dias, G., Santos, C. and Cleuziou, G. 2006. Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining. In Proceedings of the Workshop on Information Extraction Beyond the Document associated to the Joint Conference of the International Committee of Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006). Sydney, Australia, July 22.

(Dolan et al., 2004) Dolan W.B, Quirck C. and Brockett C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources, In Proceedings of 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland.

(Doucet and Ahonen-Myka, 2006) Doucet A. and Ahonen-Myka H. 2006. Probability and Expected Document Frequency of Discontinued Word Sequences, an efficient method for their exact computation. TAL journal, special issue on "Scaling of Natural Language Processing: Complexity, Algorithms and Architectures, 46 (2): 13—37.

(Dunning, 1993) Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. In *Association for Computational Linguistics*, 19(1).

(Evans and Lefferts, 1993) Evans, D. and Lefferts, R. 1993. Design and Evaluation of the CLARIT-TREC-2 System. *TREC93*. 137--150.

(Fellbaum, 1998) Fellbaum C.D. 1998. WordNet: An Electronic Lexical Database. MIT Press, New York.

(Ferragina and Gulli, 2003) Ferragina, P., Gulli, A. 2003. A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. In Proceedings of the 14th International Conference on Data Mining, San Francisco, CA, May.

(Galley and McKeown, 2003) Galley M. and McKeown K. 2003. Improving Word Sense Disambiguation in Lexical Chainin. In Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico.

(Grefenstette, 1994) Grefenstette, G. 1994. Explorations In Automatic Thesaurus scovery, Boston/Dordrecht/London, Kluwer Academic Publishers.

(Hearst, 1994) Hearst, M. 1994. Multi-Paragraph Segmentation of Expository Text, In Proceedings of the 32nd Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, June, 9--16.

(Hirst and St-Onge, 1997) Hirst G. and St-Onge D. 1997. Lexical Chains as Representation of Context for the Detection and Correction of Malapropisms. In WordNet: An electronic lexical database and some of its applications. MIT Press

(Jing and Mckeown, 2000) Jing, H. and McKeown, K. 2000. Cut and paste based summarization. In *Proceedings of NAACL*.

(Knight and Marcu, 2002) Knight K. and Marcu D. 2002. Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Inteligence, 139(1):91-107.

(Kozima, 1993) Kozima, H. 1993. Text Segmentation Based on Similarity between Words. In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session), Colombus, Ohio, USA, 286--288.

(Lin, 1998) Lin D. 1998. An Information-theoretic Definition of Similarity. In 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco.

(Liu, 2004) Liu, H. 2004. MontyLingua: An end-to-end natural language processor with common sense. Available at: web.media.mit.edu/~hugo/montylingua.

(Maarek et al., 2000) Maarek, Y., Fagin, R., Ben-Shaul, I. And Pelleg, D. 2000. Ephemeral document clustering for web applications. Technical Report RJ 10186, IBM Resarch.

(Morris and Hirst, 1991) Morris, J. and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17(1): 21--43.

(Muller et al., 1997) Muller, C., Polanco, X., Royauté, J. and Toussaint, Y. 1997. Acquisition et structuration des connaissances en corpus: éléments méthodologiques. Technical Report RR-3198, Inria, Institut National de Recherche en Informatique et en Automatique.

(Papineni et al., 2001) Papineni, K. Roukos, S. Ward, T. Zhu W.-J. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation, IBM Research Report RC22176.

(Pezner and Hearst, 2002) Pevzner, L., and Hearst, M. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. Computational Linguistics, 28 (1), March 2002. 19-36

(Phillips, 1985) Phillips, M. 1985. Aspects of Text Structure: An Investigation of the Lexical Organisation of Text, North Holland Linguistic Series, North Holland, Amsterdam.

(Ponte and Croft, 1997) Ponte J.M. and Croft W.B. 1997. Text Segmentation by Topic. In Proceedings of the 1st European Conference on Research and Advanced Technology for Digitial Libraries.120--129.

(Reynar, 1994) Reynar, J.C. 1994. An Automatic Method of Finding Topic Boundaries. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, Las Cruces, USA.

(Salton et al., 1975) Salton, G., Yang, C.S., and Yu, C.T. 1975. A theory of term importance in automatic text analysis. Amer. Soc. Inf. Sc~ 26, 1, 33--44.

(Sardinha, 2002) Sardinha, T.B. 2002. Segmenting corpora of texts. DELTA, 2002, 18(2), 273--286. ISSN 0102-4450.

(Silber and McKoy, 2002) Silber G. and K. McCoy K. 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. In Computational Linguistics, 28(4). 487--496.

(Smadja, 1993) Smadja, F. 1993. Retrieving Collocations From Text: XTRACT. In Computational Linguistics, 19 (1). 143--177.

(Turney et al., 2003) Turney P.D., Littman M.L., Bigham J. and Shnayder V. 2003. Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. In Proceedings of the International Conference on Recent Advances in Natural Language Processing.

(Veiga et al., 2004) Veiga, H., Madeira, S. and Dias, G. 2004. Webspy. Technical Report nº 1/2004. http://webspy.di.ubi.pt.

(Zeng et al., 2004) Zeng, H., He, Q., Chen, Z. and Ma, W. 2004. Learning to Cluster Web Search Results. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, UK, 210-217.