IBIMA Publishing
Communications of the IBIMA
http://www.ibimapublishing.com/journals/CIBIMA/cibima.html
Vol. 2011 (2011), Article ID 254926, 12 pages
DOI: 10.5171/2011.254926

Automatic Acquisition of Corpus for Multimedia Applications

Najeh Hajlaoui

Orange Labs, Lannion, France

Abstract

Evaluations of tools (information retrieval systems, machine learning, speech recognition, machine translation, automatic acquisition of data, etc.) are annually organized throughout evaluation campaigns (TREC, ELRA, ESTER IWSLT, etc.). The building of an *ad hoc* evaluation corpus in the context of these evaluation campaigns is a complex task and it is done manually today and with a high cost. Indeed, this is a very dedicated corpus that would answer to an application need in a precise context but automating its building is a challenge that will help significantly the organization of these campaigns. As a contribution to this challenge, we propose in a context of multimedia information retrieval, an approach of multilevel extension of a small applicative corpus to a larger and voluminous corpus based on the detection of intersections between the two corpus in terms of lemmas having the same grammatical label, that means to get a list of appropriate terminology for which we use several tools (internal and external to our laboratory) and we try to evaluate them in order to keep consistency and coherence with the original corpus..

Keywords: multimedia information retrieval, corpus for evaluation, multilevel extension, acquisition of terminology, acquisition of corpus.

Introduction

Situation

Evaluations of tools (information retrieval, machine learning, speech recognition, machine translation, automatic acquisition of data systems, etc.) are annually organized throughout evaluation campaigns (TREC, ELRA, ESTER IWSLT, etc.). These campaigns provide to participants: in a first time, a configuration corpus allowing to "optimize" the performance of each candidate system for the evaluation.

The optimisation of each tool is made in function of some functionality requested by the evaluation campaign. For example, in TREC, initially we evaluate the documents to retrieval and then we evaluate portions of documents, which are most relevant or question-answering systems (Q/R), etc.

In the earlier stage of the campaign, a second larger corpus is provided to the participants to allow them to make a final configuration and to make their system operational.

Then a set of test is provided for participants to provide in return the results of their system. To the set of test (queries type for the documents retrieval or questions type in the case of Q/R systems) correspond to a set of deliverables. The set of test and the set of deliverables constitute the evaluation corpus. This corpus is builded manually, that means with a high cost. As illustration, the INEX corpus of semi-structured XML documents is from national cooperative projects dating to 2002 until today. The most advanced studies, which are as objective to minimize the cost of building evaluation corpus, use semi-supervised methods and Rankboost algorithm to exploit directly the results of systems in competition to propose the "best" results.

Evaluation of tools requires heavy human resources and is taken now on the basis of arbitrary corpus, which does not necessarily reflect the needs under operational conditions. In addition to systems based on machine learning, operational uses require the constitution of *ad hoc* corpus, which is not available yet in the text domain like in the speech recognition domain.

In the context of multimedia applications developed in our laboratory, one of the problems to resolve concerning the improvement of research is by taking in account the terminology (phrases, named entities, etc.). Our objective is to enhance and maintain an existing base of terminologies, initially builded manually.

Our technical choice involves automatic acquisition of terminologies from learning techniques to resolve problems of quantity (completeness) and quality.

Interest and Objectives

The goal is to evaluate automatic acquisition of terminology systems, but in conditions close to the operational. The target application concerning the multimedia search VSE "Video Search Engine" but the methods and algorithms developed in this work should be generic to be reused in other research themes. For cost problems, the building of corpus must be automated to the maximum. We dispose for this a corpus of queries and a text data from the VSE application. These data are woefully insufficient from a statistical point of view and cannot serve as a training corpus.

The problem to solve is to extend from these small but available corpuses in a voluminous corpus from the Web data (collection methods, cleaning noisy corpus, errors, usage errors, mixing languages, sublanguages, sms, forum, categorization "binary" of collected corpus ...). The result of this work is to provide a common basis of learning for all tools to evaluate.

Possible Approach

To resolve this last problem which consists to build a corpus of evaluation (which can be simplified into a list of terminology) and to define objective criteria for evaluation (recall, precision, others, etc.), we can organise the solution as follows:

Expression of the Application Need: it consists to analyse the initial corpus (to extend) and is characterized by calculable criteria. For example, in VSE and for the TV subtitles, it evaluates their quality and the statistical relevance. For the purposes of enlargement to textual press corpus, it determines the adequacy of the theme VSE/press. If it is necessary to define the profile of press media to crawler, the result of this work is to build a (dynamic) "uniform" corpus allowing it to test the various tools on the same data.

Extension of Data: it consists firstly to propose one or more methods of extensions of the application corpus into a larger and voluminous corpus and ensuring the adequacy between the two corpora. Then it consists to prepare a software platform for the acquisition of terminology establishing a list of tools (those existing in our laboratory) and/or other free tools available on the Web. We deplore the available tools with the established corpus and provide updated terminology data based on the evolution of corpora. The result of this work is to build collections of data from different tools to make evaluations.

Final Evaluation: it consists to make a comparative evaluation of the obtained terminologies by the different tools used and by different methods (recall, precision, etc.). Then, it would be interesting to establish reviews and recommendations about conditions for using this type of tool with this type of training data and this type of application.

We detail in what follows, the two first steps.

Expression of the Application Need

We have in the context of the VSE project, documents indexed with the query logs and other types of news corpus. We want to build on each of these resources a learning corpus adapted to a need.

For this, we must analyse qualitatively and quantitatively our need and characterize as possible the application content.

In the following, we work on the example of the 2424actu application corpus. We begin with a detailed description of this corpus.

2424actu is a news search engine offering a multimedia content (video, audio, image, etc.) grouping and merging multiple news sources (TV, radio, press, etc.).

2424actu is a larger panorama of news. Using a simple interface, it provides broadcasts, news stories and articles, which are automatically grouped and classified by theme (international, politics, society, economy, sports, culture).

2424actu is publicly accessible under the URL **http://www.2424actu.fr/**. However, it presents a considerable interest because it represents what users search in reality.

In the context of this project, we have a certain number of accesses to the news provided by several producers in a certain form of cooperation or exchange of services. Today, the number of producers is 48.

Fig.1 shows the most important in terms of production.

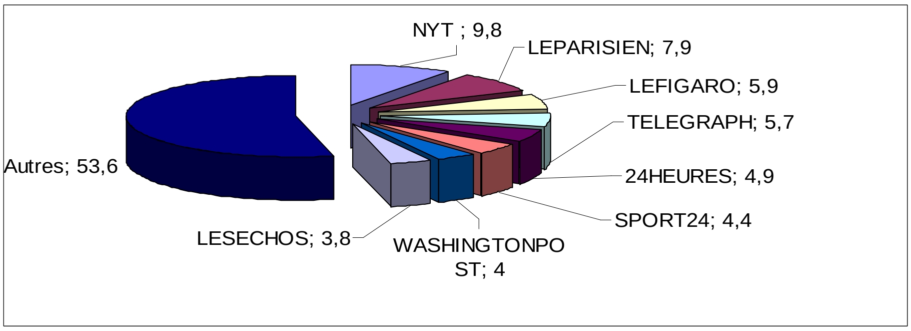

Fig.1: Principal Producers of News

Note that about 75% of this information is in French. The rest (25%) is in English because there is some news that is provided by English and American producers. We are interested in first time in processing news in French language.

We have a XML file containing news since 20/06/2009 and until today. This news is accompanied by descriptive metadata (identifier, date, producer, etc.).

In what follows, we focus on analysing the content of this XML file. In particular, we describe the process of its building and its size and its content.

News evolve every day and are updated regularly. Two ways to recover news:

We receive information in the News ML form[1].

We collect RSS news from some producers.

In both cases, if it is new news, it is registered under a new identifier in the database. If it is simply an update of old news, as detected by an identifier existing in the database (*news_id*), it is registered under the same identifier with updating the modification date.

The total corpus size is 87 megabytes; the size of the French text, composed by all the summaries of each item and located in the *<summary>* tag is 16 megabytes. The average size of content tags is 27.5 words. Naturally, there are *<summary>* tags that are empty because the news is in video or image or audio form. The size of the longer summary is 5025 bytes.

The size of the French text found in the tags *<news_title>* is 3.2 megabytes.

All text is generally clean and well-written and does not contain errors or misspellings. The most frequent words are regular words or connection words (*de, la, en, à, etc.*).

The contents of the 2424actu corpus evolutes by recovering the rest of the text using the web address found in the *<URL>* tag.

Two types of evolution can be distinguished: a static evolution and dynamics evolution.

Static Evolution: the static evolution consists to fetch the rest of the text that accompanies the information provided. For example, the text that completes the first tag *<summary>* is the following:

"*Dans cette rubrique J'ai lu, j'écoute, RFI musique vous propose d'écouter les dernières nouveautés des albums francophones dont nous avons parlé dans nos colonnes. En un clic, accédez aux extraits de celles et ceux qui font l'actualité musicale.*"
This evolves the text of this tag of 55%.

Dynamic Evolution: the dynamic evolution is linked to the contents of the *<modification_date>* tag. News may evolve having a suite or a relance such as the maritime disaster in USA.

If the content of this tag is changed, we can save the update of the news. Unfortunately, at present there is not an incremental backup of the news.

The formalisation of the need consists to normalize and to find the characterised criteria. For example, if in a corpus of queries, we constat that since multi-word sequences exist in the corpus, then these terms can be a formalization of a calculable criteria.

We formalize our need by an informational measure that gives an idea of the lexical complexity, syntactic complexity, and the richness of the vocabulary of the application corpus.

More generically, we want to have a most voluminous and larger corpus under a number of consistency and coherence constraints. We then try to detail the concept of the corpus and its characteristics. We note here the principals notions linked to corpus concept.

Nelson, F. W. (1982), defines corpus as:

« A collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis ».

A corpus is considered as a set of documents (texts, images, videos, etc.) regrouped in a precise optic. We can use corpora in several domains: literary studies, linguistic, scientific, etc.

[1] News ML is a specific format for news

In literature, a corpus is a collection of texts with a common purpose. In science, the corpora are essential tools and valuable in natural language processing. They allow extracting a set of useful information for statistical treatment.

From an informative viewpoint, they allow building sets of frequencies of n-grams. From a methodological point of view, they allow necessary objectivity for scientific validation in natural language processing. Information is not empirical; it is verified by the corpus. It is therefore possible to use corpora to generate and verify scientific hypotheses.

Several characteristics are important to create a well-formed corpus such as:

The Size: corpus must obviously reach a critical size to allow reliable statistical treatment. It is impossible to extract reliable information from a too small corpus.

The Language Corpus: a well-formed corpus must necessarily cover a single language, and one variation of that language. For example, there are subtle differences between the French of France and the French spoken in Belgium. It is therefore not possible to derive reliable conclusions from a Franco-Belgian corpus for French in France or for French in Belgium.

The Evolution of the Texts over Time: time has an important role in the evolution of language: the French spoken today is not the French spoken 200 years ago or, more subtly, the French spoken 10 years ago, especially because of neologisms. It is a phenomenon to take all languages into account. A corpus should not contain texts written at too long time intervals.

The Register: do not mix different registers, a corpus builded from scientific texts cannot be used to extract information from vulgarised texts and a corpus of scientific and vulgarised texts will not allow any conclusion on these two registers.

In this work, we try to build a larger and voluminous corpus by the extension of a smaller corpus and respecting the previous characteristics.

Extension of Corpus

We did not find previous work about extension of application corpus for an objective of extraction and enrichment of terminology. There is other work but in a different context such as JRC team of the European Commission, which work on the calculation of similarity between multilingual documents using as pivot the EUROVOC as in research by Steinberger, R. Pouliquen, et al (2002). In this context, the hybrid approach based on a combination of TTR, likehood, Okapi, distance calculation methods has shown its effectiveness. More details about Okapi are in Robertson, S. E. et al (1994).

Lafourcade M. et al (2009) has made a web-based game to collect terms by building a lexical network. Their approach consists in having people take part in a collective project by offering them a playful application accessible on the web. From an already existing base of terms, the players themselves thus build the lexical network, by supplying associations, which are validated only by an agreeing pair of users. These typed relations are weighted according to the number of pairs of users who provide them. This game has now about 180,000 relations.

Here, we are interested in finding a solution for extending an existing corpus to a larger and voluminous corpus keeping adequacy. We find several problems.

The First Problem consists in the matching between two structured or unstructured documents d and D where d is a document from the application corpus and D is a larger document from the corpus automatically acquired, a problem of different logical structures for the pair (d, D), a problem of likelihood of their logical structures, a problem of likelihood of their content, etc.

The Second Problem is algorithmic problem; it consists to know how to cross n (thousands) documents of the application corpus with a few m (millions) extensive corpus of documents D.

The Third Problem is to know how to clean effectively the extended corpus to optimise the adequacy function with the application corpus.

Process of Extension

Two cases are presented to expand the existing application corpus: an extension from the same corpus saving some correspondence (alignment in logical structure level) and an extension from another larger corpus without any information about correspondence.

The first case consists to enrich the application corpus with query results formed from the corpus itself and with these query, we search an equivalent but most larger corpus (from structure point of view). For example, we identify for each part of the application corpus, the most frequent terms (compound words, multi-words, etc.) and pass them as queries.

In the case of our corpus 2424actu, we can get the rest of the data through the URL provided with news.

In the case of an extension of the application corpus to a most larger and voluminous corpus with an equivalent structure, we propose an approach, which measures the variation of vocabulary and it is based on a modification of the measure TTR (Type Token Ratio).

We called this method LTTR (Local Type Token Ratio) that is calculated locally for each text portion of the document d. In the case of the 2424actu corpus, we calculate the LTTR for each tag content <*summary*>.

In the case of an extension of the application corpus from another foreign corpus without an equivalent structure, we propose a multilevel extension approach. Really, we have a corpus from the Web (2 G.O) and we try to find an approach to get adequate text from the big corpus and to add it to the original corpus.

We present now a generic approach which can be used in the too cases (with and without correspondence). Note that here in our case, the two corpus are composed respectively of one document d and D.

Extension Level 1: as shown in Fig. 2, we start with an operation of lemmatisation and grammatical tagging, after which, the two documents lose their logical structure, keeping the history and traces of the origin of each term. Then, we detect all lemmas noted *types(d, D)* which is the intersection of d and D (intersection on lemma and grammatical tagging). For each $lemma_i$ of the set *types(d,D)*, we will search all terms noted *Terms1(D)* and containing the terms T_1 T_2 ... T_m of document D and converging to the $lemma_i$

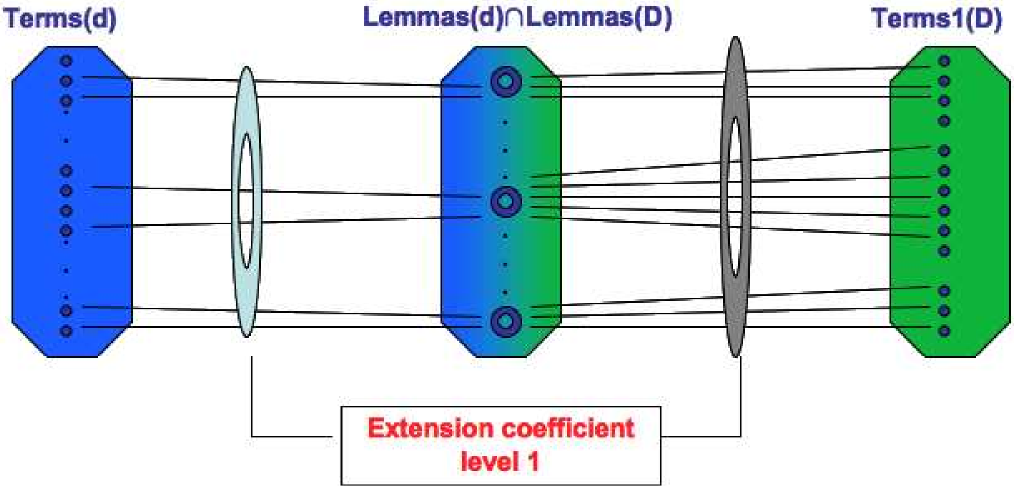

Fig. 2 : Extension Level 1

We add new terms introduced from documents D to the document d and we define the extension coefficient level 1, the ratio of the sum of terms of D noted *Terms1(D)* which converge to the lemmas of intersection by the sum of terms of d that converge to the same intersection lemmas.

Extension Level 2: as shown in Fig. 3, we apply the same steps of the extension level 1 and we pass an extension level 2 by getting the texts that correspond to each term T_1 T_2 ... T_m of document D.

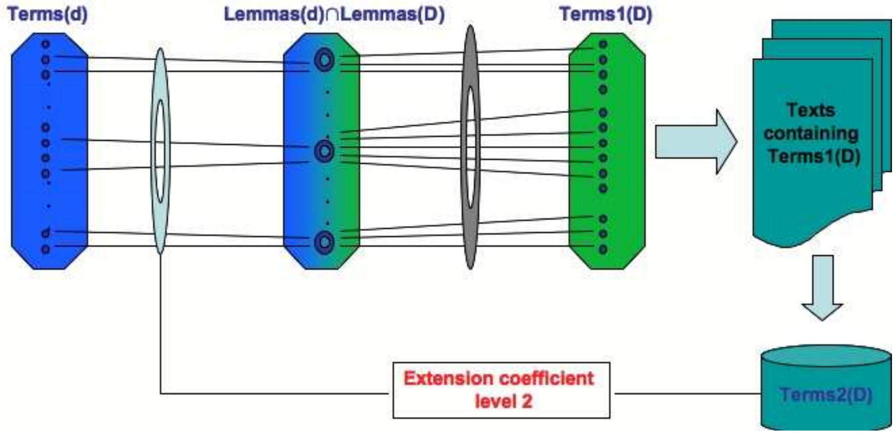

Fig. 3 : Extension Level 2

Theoretically, all the texts founded we produce a set of terms *Terms2 (D)* larger and more voluminous than *Terms1 (D)* which satisfies the constraint extension level 1.

Similarly, we define the extension coefficient level 2, the ratio of the sum terms *Terms2 (D)* by the sum of terms of d that converge towards the same intersection lemmas.

Extension Level 3: we can get a larger extension of the initial corpus making two types of semantic rapprochement as shown in Fig.4:

Direct rapprochement: it consists to regroup some terms of D that are not in the intersection set *types (d,D)* to certain terms of d.

Indirect rapprochement: it consists to regroup some terms of D that are not in the intersection set *types (d,D)* to certain terms of D which are in *types (d,D)*.

Then we continue the same process of extension level 2. This gives us a set of terms larger and more voluminous noted *Terms3 (D)* for which we define in the same way an extension coefficient level 3.

Note that the rapprochement can be at the semantic level. For example, regroupping the term "grippe A" to the term "grippe Z" or "réchauffement du climat" to "réchauffement climatique".

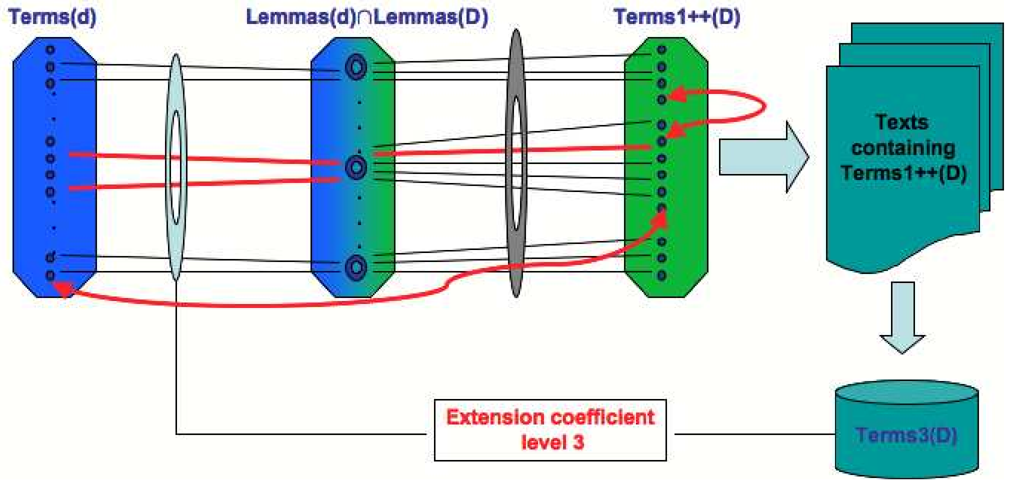

Fig.4: Extension Level 3

This approach of multilevel extension can be applied to the first case of the extension with correspondence.

Evaluation and Matching Approach

We evaluate the result of the acquisition of terminologies by classical methods such as recalling and precision, which can be based on a measurement of terminology distance (Nazarenko et al., 2009).

We define the formula (1) which allows us to calculate terminology distance $D_{termino}$ from the number of adequate terms k in the candidate corpus. The formula (2) allow de calculate ($D_{termino}$ in function of k).

In this work, we use formula (1). That means, we start by calculating the number of adequate terms having the same lemma and the same grammatical labels. If not, we limit to have the same lemmatization. We calculate after a F-measure that takes into account the length of the two corpus (reference and candidate).

The following are two examples of adequate terms: *internationaux* and *morte* are respectively adequate to *internationales* and *mort*.

internationales [ADJ. international], internationaux [ADJ. international]

mort [NOM. mort], morte [ADJ. mort]

We suppose:

Ref: a reference corpus, and |Ref|=m

Cand: a candidate corpus, and |Cand|=n

K = |{adequate terms in the candidate corpus }|

$D_{termino}(Ref,Cand)=m+n-2K$ (1)

$K = (m+n- D_{termino}(Ref,Cand))/2$ (2)

We calculate the recall R, precision P, and F-mesure F):

Recall: $R=K/m$

Precision: $P=K/n$,

F-measure: $F=1- D_{termino}(Ref,Cand)/(m+n)$

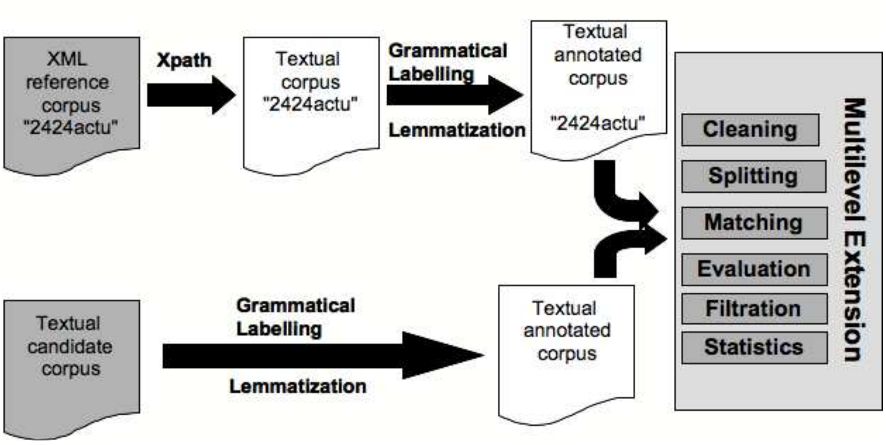

Fig. 5: Architecture of the Multilevel Extension

We split the two corpora (candidate and reference) in small parts with parameters of size respectively equal to m and n. Fig.5 describes the experimental architecture. The multilevel extension takes as input two annotated textual files (candidate and reference). Until a fixed F-measure (F_{limit} = 0.5), we decide to reject or accept the new part coming from the candidate corpus.

Results

We developed the two first levels of extensions. We experimented the extension operation by varying the sizes of the reference and candidate corpus as well as the sizes of block (m and n). Table 1shows some maximum values of F-mesure (F) and the ability to discover new terms refers to the two extension levels (level 1 (L1) and level 2 (L2)).

Table 1 : F-Mesures and % of Extension

		Reference corpus			
		Size	1000(747 T/ 368 T≠/ 322 L≠)	10000(784 T/ 2916 T≠/ 2453 L≠)	100000(77910 T/12742 T≠/ 9576 L≠)
Candidate corpus	Size	Splitting	100	10000	10000
	1000(766 T/ 446 T≠/ 415 L≠)	100	F=0,53 N1=3% N2=68%	$F<F_{limit}$	$F<F_{limit}$
	10000(7379 T/ 2706 T≠/ 2341 L≠)	10000	$F<F_{limit}$	F=0,74 N1=13% N2=72%	$F<F_{limit}$
	100000(79373 T/ 10666 T≠/ 7444 L≠)	10000	$F<F_{limit}$	$F<F_{limit}$	F=0,75 N1=14% N2=54%

We can read Table 1 as follows: for a reference corpus composed by 1000 terms (equivalent to 747 terms "T" after cleaning, and 368 different terms "T ≠" and 322 different Lemmas "≠ L") and a candidate corpus of 1000 terms (equivalent to 766 terms "T" after cleaning, and 446 different terms "T ≠" and 415 different Lemmas "≠ L"), we obtain a F-measure equal to 0.53 and a percentage discovering new terms for level 1 equal to 3% and for level 2 equal to 68%.

The experiments consist to detect a maximum F-measure and therefore a maximum percentage of extension for each pair of corpus trying several parameters of splitting (100, 1000, 10000, etc.) depending on the size of corpus. Fig. 6 shows the different values of F-measure of Table 1 for sizes equal to 1000, 10 000 and 100 000 terms. Obviously, we obtain the same experimental values of F-measure for both levels of extension. Thus, the two curves overlap.

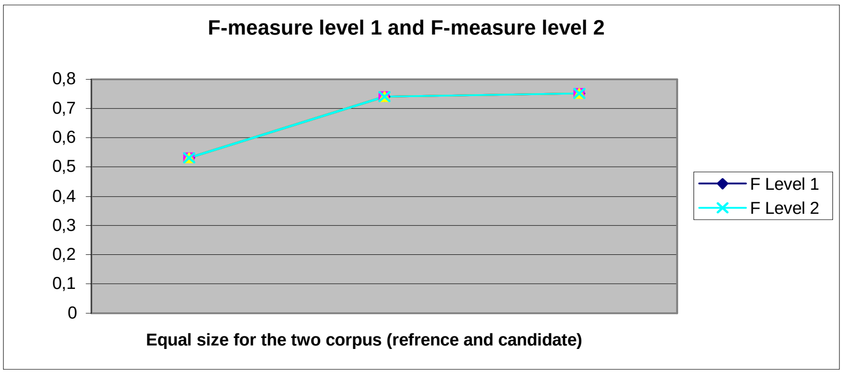

Fig.6: F-measure Level 1 = F-measure Level 2

Fig. 7 shows an example of the number of new terms added to the reference corpus. For example, we extended the 368 different terms of the reference corpus to 379 different terms using the extension level 1 and to 618 different terms using the extension level 2.

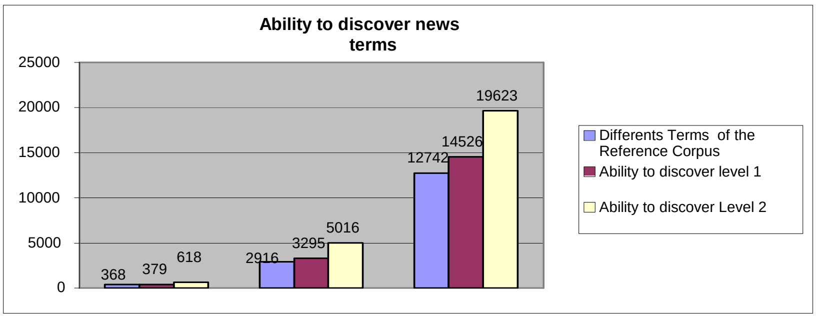

Fig.7: Example of Ability to Discover News Terms

Conclusion

We have presented the extension problem of an application corpus to a larger and voluminous corpus which is the first step to acquire a list of appropriate terminology.

We have analysed this problem by showing that there are two different cases of extension: an extension with or without logical structure correspondence.

We have proposed a generic method that can be applied in both cases. It consists on a multilevel extension of a small application corpus from a larger corpus based on the calculation of intersection of the two corpus having the same lemmatisation and grammatical tagging. Hence having a good result of lemmatisation and grammatical tagging is very important.

We experimented the two levels of extension and we got good results of extension, which allow us in the future to experiment with larger data.

The advantage of this approach is that it is multilevel and multilingual. Indeed, it can be applied to languages other than French. It provides a configuration for quality and / or quantity of the new data by adjusting the size of the paramerters to split initial corpus (in small or large blocks).

References

Baroni, M. & Bernardini, S. (2004). "BootCaT: Bootstrapping Corpora and Terms from the Web," Proceedings of LREC 2004.

Baroni, M. & Ueyama, M. (2004). "Retrieving Japanese Specialized Terms and Corpora from The World Wide Web," Proceedings of KONVENS 2004.

Collins, M. (2000). 'Discriminative Reranking for Natural Language Parsing,' In Proceedings of the Seventeenth International Conference on Machine Learning.

Freund, Y., Iyer, R. D., Schapire, R. E. & Singer, Y. (2003). "An Efficient Boosting Algorithm for Combining Preferences," *Journal of Machine Learning Researc*h 4. Pages 933-969.

Freund, Y. & Schapire, R. E. (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55(1):119139, A

Géry, M., Largeron, C. & Thollard, F. (2009). "Impact Précoce du Poids des Balises pour la

Recherché D'information Ciblée," CORIA 2009. 5-7 mai. Toulon.

Hajlaoui, N. (2008). 'Multilinguïsation de Systèmes de E-Commerce Traitant des Énoncés Spontanés en Langue Naturelle,' *Thèse*. Université Joseph Fourier. Grenoble. 25 septembre 2008. 318 p.

Iyer, R. D., Lewis, D. D., Schapire, R. E., Singer, Y. & Singhal, A. (2000). "Boosting for Document Routing," In Proceedings of the Ninth International Conference on Information and Knowledge Management.

Lafourcade M. & Joubert A. (2009). "Similitude Entre les sens d'usage d'un Terme Dans un Réseau Lexical," *Revue TALN*. Volume 50-n°1/, pages 177-200.

Nazarenko, A., Zargayouna, H., Hamon, O. & Puymbrouck, J. V. (2009). "Évaluation des Outils Terminologiques: Enjeux, Difficultés et Propositions," *Revue TALN*. Volume 50-n°1/, pages 257-281.

Nelson, F. W. (1982). 'Problems of Assembling and Computerizing Large Corpora,' Proc. Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities. pp. 7-42.

Nelson, F. W. (1992). 'Language Corpora B.C. Proc. Directions in Corpus Linguistics,' Proceedings of Nobel Symposium. Stockholm. pp. 17-32.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M. & Gatford, M. (1994). 'Okapi in TREC-3, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA,' NIST Special Publication 500-225, pp. 109-126.

Sharoff, S. (2006). "Creating General-Purpose Corpora Using Automated Search Engine Queries," In Baroni and Bernardini (eds.) Wacky! Working papers on the Web as Corpus. Bologna: GEDIT.

Steinberger, R. Pouliquen, B. & Hagman, J. (2002). "Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC," Third International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2002, Mexico City, Mexico, 17-23 February

Vu, H.-T. & Gallinari, P. (2006). "Apprentissage Statistique pour la Constitution de Corpus d'évaluation," CORIA 2006, Lyon, France, 15-17.

Walker, M. A., Rambow, O. & Rogati, M. (2001). "SPoT: A Trainable Sentence Planner," In Proceedings of the 2nd Annual Meeting of the North American Chapter of the Associataion for Computational Linguistics.