A Xml-Based Term Extraction Tool for Basque

I. Alegria (1), A. Gurrutxaga (2), P. Lizaso (2), X. Saralegi (2), S. Ugartetxea (2), R. Urizar (1)

(1) Ixa taldea –University of the Basque Country
649 Postakutxa. 20080 Donostia
acpalloi@si.ehu.es
(2) Elhuyar Fundazioa
Astesuain Poligonoa, 14 - 20170 Usurbil
agurrutxaga@elhuyar.com

Abstract

This project combines linguistic and statistical information to develop a term extraction tool for Basque. Being Basque an agglutinative and highly inflected language, the treatment of morphosyntactic information is vital. In addition, due to late unification process of the language, texts present more elevated term dispersion than in a highly normalized language. The result is a semi-automatic terminology extraction tool based on XML, for its use in technical and scientific information managing.

1. Introduction

In the last years, an increasing interest in term extraction is reflected in several works. Moreover, different research groups have developed several tools for automatic terminology extraction from specialized corpora: ACABIT, ANA, LEXTER, TERMINO, TERMS, Xtract, CLARIT, FASTR, NODALIDA…

The aim of this project, based on linguistic and statistical techniques, is to develop a term extraction tool for Basque. Being Basque an agglutinative and highly inflected language, treatment of morphosyntactic information is vital. In addition, due to late unification process of the language, texts present more elevated term dispersion than in highly normalized languages.

The result is a semiautomatic terminology extraction tool based on XML, for its use in technical and scientific information managing. Due to the modular architecture of the system, the future multilingual capability of the application is also considered. The different modules exchange and manage information in XML.

The specialized corpus used for testing the tool consists of 2,706,809 words and includes all the articles published by Elhuyar Foundation in the zientzia.net site (www.zientzia.net) until 2003. A sample of 28 documents with 13,756 words and composed exclusively by computer science divulgation articles has been processed manually. The articles of this corpus are very short, so the expected results may be worse than other ones in the literature. In the second phase of the project, a larger hand-tagged corpus will be available after a semiautomatic process.

2. Linguistic Process

Most systems preprocess texts in order to obtain POS tags of words. After this step, some tools use only statistical techniques, identifying Noun Phrases (NP) or other syntactic patterns. In some systems semantic information is used combined with statistical information when ranking and clustering of candidates is carried out.

The aim of the project is to use a linguistic module in order to obtain a list of candidates with a very high recall and to let the statistical module rank these candidate-terms.

In our system, only Noun Phrases are considered for the selection of the terminological units. In order to select the most usual NP structures of Basque terms, we have taken as a starting point the previous work done by the IXA group (Urizar et al. 2000). Moreover, a new study on the morphosyntactic structures of terms has been carried out using the reference sample. As a result, we have added some new patterns in order to increase recall.

The next step applies *Euslem* (Ezeiza et al., 1997) for lemmatization and POS tagging, and a grammar identifies the terms corresponding to the structures of most representative and productive patterns. Those patterns stem from the same patterns of the manual extraction.

The Grammar

In a grammar defined using xfst (Beesley & Karttunen, 2003), the tokens of the patterns are described according to the output of the tagged corpus. The combination of those tokens defines the patterns in the grammar. The grammar is written in xfst syntax and then compiled on a transducer which reads the tagged corpus in order to extract the maximal NPs that match the patterns. The mentioned grammar also deals with the position of numbers, typographical elements such as capitalization, hyphen insertion and so on in specific patterns and foreign words that are not recognized by the tagger.

Term Normalization and Nested Terms

The candidate terms are retrieved in their canonical form, so basic normalization is carried out. For instance, *sistema eragile* ('operating system') is obtained from all its different inflections: *sistema eragileari* (dative sing.), *sistema eragilearen* (genitive sing.), and so on.

As we said before, most of the typographical variations as capitalization, hyphen insertion, etc. are treated too. Syntactic and semantic variations are not managed by the moment.

In order to deal with terms included in longer combinations (nested terms), the linguistic module decomposes those maximal NPs into sub-structures. In order to obtain these syntactic constituents (*head* and *modifier*) a very simple grammar has been written based on the probability of the components in each pattern. In

LEXTER (Bourigault, 1994), an automatic extraction module is used for this proposal.
For example, a sentence in the grammar states that in the pattern N-N-Apos[1] N-N remains as possible term. For instance, from *RAM memoria handi* ('big RAM memory') *RAM memoria* is added to the list of candidate terms. This process is very important since the main aim of this module is to achieve a high recall. Tables 1 and 2 show the recall improvement obtained when the treatment of nested terms is carried out. Recall grows from 66% to 87%.

Type	# of terms	# of candi-dates	# of correct terms	recall (%)
One-word units	245	690	158	64
Bigrams	255	859	170	67
3-4-grams	60	341	41	68
TOTAL	560	1890	369	66

Table 1. Results without treatment of nested terms

Type	# of terms	# of candi-dates	# of correct terms	recall (%)
One-word terms	245	1179	238	97
Bigrams	255	1156	210	82
3-4-grams	60	341	41	68
TOTAL	560	2676	489	87

Table 2. Results with treatment of nested terms

Some terms are not identified as candidates mainly because errors are produced in the tagging and lemmatization process. After this linguistic process, a list of term candidates is generated. The statistical module will classify the terms in this list and apply a threshold, which will vary depending on the application.

3. Statistical Process

The statistical methods applied in this kind of applications vary considerably depending on the system. In our approach, we decided to apply two different strategies for multiword and for one-word terms. *Unithood* is used by means of word association measures for the treatment of multi-word candidates, and *termhood* measures (Kageura, 1996) for one-word term candidates. The association measures offer good results when the frequency of the terms is not too low; however, the results decline when terms occur once (Hapax Legomena).
In our corpus, we encountered some problems when applying association measures due to the high presence of Hapax Legomena (table 3) and the lack of representativeness of word frequency.

Type	freq terms =1	freq terms =2	freq terms >2
One-word terms	28.20%	16.00%	55.80%
Multiword terms	77.47%	12.91%	9.62%

Table 3. Terms frequency

[1] *N*: non case noun, *Apos*: postpositive adjective

In our experiments, the association measures were empirically modified trying to introduce a simple *termhood* paradigm. In this case, the changes improve the ranks.

Multiword Terms

Word association measures are used in order to rank multiword units according to the association grade among their components. This association grade or *unithood* would determine their *termhood* for any domain. For single domain corpus, it is a good *termhood* approximation if data dispersion is not high.
Most of the association measures proposed in the literature are intended to rank bigrams and relay on different concepts. For example, Mutual Information (MI), introduced in this field by Church and Hanks (1989), was taken from Information Theory. Other measures such as the log-likelihood ratio (LL) introduced by Dunning (1994), t-score and Chi-square are based on hypothesis testing.
In order to rank MWUs composed of two or more words, Dias et al (2000) introduced Mutual Expectation (ME), a measure based on Normal Expectation, which is a generalization of Dice coefficient for N-grams. Blaheta and Jonhson (2001) use measures based on parameters of certain Log-linear models to rank verbs composed of two or more words.
Other measures incorporate linguistic information. C-NC-SNC-Values (Maynard and Ananiadou, 2000) measure the independence grade of the multi-word-units in the corpus and use semantic knowledge and context information. Nakagawa (2003) takes into account the fact that complex terms may be composed of simple terms, and describes measures that ponder these relations. Lapata and Lascarides (2003) combine the probabilities of candidate constituents to be good term components, the semantic relations between the components and Machine Learning based classifiers to evaluate those features.

Testing bigrams

The input is the list of candidates extracted in the linguistic process. We have carried out experiments with two lists: with and without processing nested terms, in order to find out the best starting point to get maximum precision and recall.
We did some preliminary experiments using LL, MI, MI^3, t-score, Chi-square, ME and measures based on Log-linear models proposed by Blaheta (the last two only with candidates of more than 2 words).
The ranks obtained with n-grams ($2 \leq n \leq 4$) were quite poor. This might be due to the characteristics of our corpus which is small and made up of short documents dealing technical topics but at divulgation level. Therefore, the high dispersion among the words and terms, the great amount of Hapax Legomena, and the lack of representativeness of word frequency make it difficult to measure the association level.
The results for bigrams using different association measures are shown in Table 3 (number of terms, precision, recall and F-score). Due to the type of corpus, the results are very similar for all the measures.

Type	# of terms	# of extr. terms	# of correct terms	P (%)	R (%)	F (%)
MI	255	1156	210	18.17	82.35	29.77
MI^3	255	1156	210	18.17	82.35	29.77
LL	255	612	135	22.06	52.94	31.14
t-score	255	681	143	20.99	56.08	30.55
Chi-square	255	1156	210	18.17	82.35	29.77

Table 3. First results for bigrams

We tried to improve empirically association measures for the characteristics our corpus. In order to improve the representativeness of word frequency and to weigh up positively candidates composed of one-word terms, when LR, MI, MI^3, t-score and Chi-square are calculated, we use for the frequency (marginal frequency) of the components their *normal frequency* instead of the observed frequency in the corpus. This normal frequency is calculated from a global character corpus. In this way, the frequency of common words is more representative and the frequency of one-word terms is smaller, increasing the ratio of the candidates composed of one-word terms.

As Table 4 shows, results improve considerable, almost 10% in F-score for all the measures. Results change if we also take nested noun phrases (second list) as candidates (Table 4). Although the results are similar in F-score measure, if we are interested in high recall, results are better when we consider nested noun phrases.

Type	P	R	F
MI	30.91	66.67	42.24
MI^3	33.26	58.04	42.29
LL	30.55	65.88	41.74
t-score	31.11	60.39	41.07
Chi-square	28.76	69.02	40.60

Table 4: Precision, recall and F-score (nested included)

These results (table 4) may vary depending on the patterns (see Table 5).

Pattern	# of terms	# of extracted terms	# of correct terms	F-score
N-N	138	269	94	46.19
N-Apos	63	116	42	46.93
Aprep-N	54	98	25	32.89

Table 5: Results depending on the patterns with MI

Experimentation with Longer Terms

We have followed two strategies to rank trigram and tetragram candidates. In those strategies, candidates are ranked by their *unithood*, but this *unithood* is estimated between different groups of constituents. In the first strategy, it is calculated between the head and modifier of the candidate (see the extraction of constituents in nested terms). In the second strategy, it is calculated among all the components.

We have observed that, as in the case of the bigrams, the classification improves when normal frequencies of the monograms are taken into account. This has been applied to all measures except for ME, where frequencies of monograms are not used. The results are showed in table 6. The two last rows (ME and measures based on Log-linear models) are calculated using the second strategy. LL, T-Score and ME measures show the best performance.

Type	P (%)	R (%)	F (%)
MI	20.69	40.00	27.27
MI^3	20.93	45.00	28.57
LL	21.71	46.67	29.63
t-score	21.71	46.67	29.63
Chi-square	20.69	40.00	27.27
ME	27.03	33.33	29.85
Log-Linear models	18.14	61.67	28.03

Table 6. Precision, recall and F-score

One-Word Terms

There are several methods to obtain the *termhood* of a one-word candidate. For example, tf-idf, which is used regularly, compares the frequency of a given word in the document with its frequency in other documents of the corpus. Matsuo and Ishizuka (2002) use statistical information about word co-occurrence.

Since we had no adequate document set available to test the tf-idf measurement, we considered that the relation between the relative frequency of the nouns and the relative frequency of a general corpus (*normal frequency*) might be a good measure to classify individual word candidates. This way, the *termhood* of the candidates is obtained dividing the observed frequency in the corpus by the normal frequency. Damerau (1993) defines this as relative frequency ratio (RFR). When RFR is applied, the best F-score occurs when the nested candidates are included (see results in Table 7).

Type	# of terms	# of extracted terms	# of correct terms
Monograms RFR	245	407	153
	P (%)	**R (%)**	**F (%)**
	37.59	62.45	46.93

Table 7: Precision, recall and F-score (nested included)

4. The Tool

Since extracting terminology is the main aim of our project, a tool was designed and implemented for this purpose. The tool is composed of the following main elements: the corpora builder, the terminology tagger, and the corpora navigator.

The corpora builder offers the user the possibility of creating a corpus from text documents of different formats and structures. Then the module converts them into an intermediary format (XML) and structure (TEI). This step helps to integrate the terminology tagger and navigation over the results in the original structure of the document.

The terminology tagger finds candidate terms in the created corpus using the methods described in previous sections. The result is the corpus with candidate terms and associated statistical information tagged using XML.

The corpora navigator offers the user the possibility of navigating through the results in their context (see Fig. 1). The original elements of the documents will be accurately displayed. An additional advantage of the format is that queries based on document structure can be carried out using X-Path.

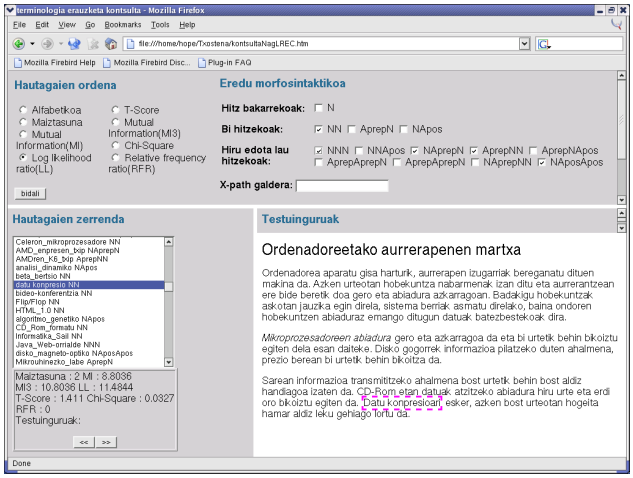

Figure 1: Interface of the application

As three-tier logical architecture has been used, user interface, process logic and data management have been put separately. The physical design lies on a web browser, a web server (Apache+mod_perl) and a native XML database (Berkeley DB XML).

5. Conclusions and Current Work

We have presented a modular tool for extracting terminology of Basque that is based in language engineering, statistics and XML. State of art in this area shows that both good recall and good precision are not possible. As we have mentioned above, the characteristics of the test-corpus limit the consistency of the system. Therefore, we are compiling a bigger test-corpus to gain credibility in our experiments.

According to the results, the main sources of errors are problems to identify foreign words and postpositions. The tagger is being improved in order to manage foreign words more efficiently. Besides, most postpositions in Basque are analysed as nouns and, therefore, they produce a non-negligible amount of noise. In order to avoid it, new rules to treat postpositions are being developped so as to be inserted in the tool. Moreover, the treatment of morphosyntactic and syntactic term variation will be taken into account in future developments of the tool.

Finally, we are planning to improve those results in the second phase of the project, where after improving some elements of the first versions, machine-learning paradigm and semantics will be used.

6. Acknowledgements

This project is part of the Hizking21 project (www.hizking21.org), which is being developed to support and achieve technical and scientific aims in the field of language technologies. This research has been partially funded by the *Etortek* and *Saiotek* programs of the Department of Industry, Trade and Tourism of the Government of the Basque Country.

7. Bibliography

Beesley, K.R. & Karttunen, L. (2003). Finite State Morphology. CSLI. Standford University.

Blaheta, D. & Johnson, M. (2001). Unsupervised learning of multi-word verbs. In Proceedings of the 39th Annual Meeting of the ACL (pp. 54-60). Toulouse.

Bourigault, D. (1994). LEXTER, un Logiciel d'Extraction de Terminologie. Application a l'acquisition des connais-sances a partir de textes. Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Church, K.W. & Hanks, PP. (1989). Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Meeting of the ACL (pp. 76-83). Vancouver.

Daille, B. (1995). Combined approach for terminology extraction: lexical statistics and linguistic filtering. In UCREL Technical Papers, 5, University of Lancaster.

Damerau, F.J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. In Information Processing & Management, 29, 433–447. Elsevier

Dias, G., Guilloré, S. Bassano, J.C. & Lopes, J.G.PP. (2000). Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? In Proceedings of Recherche d'Informations Assistée par Ordinateur (pp. 1-20). Paris.

Dunning, T. (1994) Accurate Methods for the Statistics of Surprise and Coincidence. In Computational Linguistics 19(1): 61-74. Cambrigde, Mass: The MIT Press.

Evert, S. & Krenn., B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In Proceedings of the 39th Annual Meeting of the ACL (pp. 188-195).

Ezeiza N., Aduriz I., Alegria I., Arriola J.M. & Urizar R. (1998). Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In COLING-ACL'98, Montreal.

Jacquemin, C. (2001). Spotting and Discovering Terms through Natural Language Processing. Cambridge, Mass.: The MIT Press.

Kageura, K. & Umino, B. (1996). Methods of Automatic Term Recognition. In Terminology. 3(2), 259-289. Amsterdan: John Benjamins.

Lapata, M. & Lascarides. A. (2003) Detecting Novel Compounds: The Role of Distributional Evidence. In Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics (pp. 235-242). Budapest.

Matsuo, Y. & Ishizuka, M. (2000). Keyword extraction from a document using word co-occurrence statistical information. In Transactions of the Japanese Society for Artificial Intelligence, 17(3), 217-223.

Maynard, D. & Ananiadou, S. (2000), Trucks: A Model for Automatic Multi-Word Term Recognition. In Journal of Natural Language Processing, 8(1), 101-126.

Nakagawa, H. & Mori, T. (2003). Automatic Term Recognition based on Statistics of Compound Nouns and their Components. In Terminology, 9(2), 201-219. Amsterdan: John Benjamins.

Urizar R., Ezeiza N. & Alegria I. (2000). Morphosyntactic structure of terms in Basque for automatic terminology extraction. In Proceedings of the 9th EURALEX International Congress. (pp. 373-382). Stuttgart.