

ELeXBI, A BASIC TOOL FOR BILINGUAL TERM EXTRACTION FROM SPANISH-BASQUE PARALLEL CORPORA

A. Gurrutxaga, X. Saralegi, S. Ugartetxea

Elhuyar Foundation

Iñaki Alegria

IXA group – University of the Basque Country

ABSTRACT

We present the work done by Elhuyar Foundation in the field of bilingual terminology extraction. The aim of this work is to develop some techniques for the automatic extraction of pairs of equivalent terms from Spanish-Basque translation memories, and to implement those techniques in a prototype. Our approach is based on a previous monolingual extraction of term candidates in each language, the creation of candidate bigrams from both segments of the same translation unit, and, finally, the selection of the most likely pair of candidates, based mainly on statistical information (association measures) and cognates. In the first step, we use linguistic techniques for the extraction of term candidates. The result of our work is ELeXBI, a prototype tool that can extract equivalent terms from Spanish-Basque translation memories. This work wants to be a contribution to corpus-based bilingual lexicography and terminology in Basque.

ELexBI, a Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora

1 Objective

The aim of this work is to develop and apply techniques for the automatic extraction of pairs of equivalent terms from Spanish-Basque translation memories, and to implement those techniques in a user-friendly prototype. This work can be located in a wider research area. In fact, the extraction of equivalent terms from translation memories can be seen as a particular case of the extraction of lexical equivalences from parallel corpora.

In this first stage of development, the translation memories that we use as input are the product of translators' work; that is to say, the alignment at sentence level is 100% correct. In future work, automatically aligned memories and parallel corpora aligned at document level will be used. As for the type of term equivalents we attempt to find, we deal with one-word and multiword terms which have noun-phrase structure. Furthermore, equivalences between one-word and multiword terms are also taken into account.

2 Extraction process

Different approaches have been proposed for the extraction of lexical correspondences from parallel corpora. Most of them are closely related to the task of word-level alignment. According to Tiedemann (2003) and Kraif (2002a, 2002b), we look at the extraction of lexical correspondences as a different but much related task.

The extraction strategy relies on the hypothesis that a given term has a single translation per corpus (Fung, 1998). As Somers points out (Somers, 2001), this condition is hardly fulfilled in real texts, but usually it is a good approach for many applications, specially for specialized domains. This hypothesis improves the precision of the extraction process because it helps blocking the indirect association that may occur. On the other hand, in a text where term variation is high the recall might be poorer. We will describe now the main steps of the methodology implemented.

2.1 Monolingual extraction of term candidates for each language

Extraction of Basque candidate terms is carried out by *Erauzterm*, a tool developed by Elhuyar and IXA (Alegria et al., 2004a, 2004b). *Erauzterm* uses linguistic and statistical techniques, and extracts NP structure candidates in their canonical (unflexioned) form, along with their morphosyntactic pattern, statistical measures and contexts. For the extraction of Spanish candidates, we use *Freeling 2.1*, an open source suite of language analyzers (Carreras et al., 2004). We chose an

output based on a tree-structure of the sentence (Shallow Parsing), and we take as candidates all the noun phrases from that tree as candidates (<grup-nom>). A problem we have identified is that only some prepositional phrases (<sp>) are attached to the NPs, while the rest are discarded. PP attachment is necessary to detect Spanish terms like *televisión por cable* or *conexión a Internet*. We have therefore attached all the <sp> elements to the preceding NP (even though we are aware of the amount of noise generated at this step).

As for the treatment of nested terms, *Erauzterm* decomposes maximal NPs into head and modifier, and selects them as candidates if they match a given pattern of the grammar. In the case of *Freeling*, we take the NPs that appear embedded in other NPs as candidate too. It has to be pointed out that a little imbalance occurs between monolingual extractions when bilingual candidates are generated, due to the different extraction criteria used by term-extractors.

2.2 Generation of bilingual candidates

In this step of the process we combine term candidates from both languages present in the same translation unit. The result is a set of pairs of term candidates or 'bigrams'. Those bigrams are stored in a relational database which includes, among others, data about one-to-one candidate segments, LCSR, AM values... Nevertheless, at this point we don't calculate LCSR and AM values. Those calculations are made within the corresponding step of selection process that we will explain in the next section.

2.3 Selection of equivalents

From the whole set of possible equivalent pairs, the selection algorithm chooses the 'best' ones according to the order of the steps of the extraction algorithm. This algorithm is based on *competitive linking* (Melamed, 2001) or *meilleur affectation biunivoque* (Kraif, 2002b). By means of an incremental algorithm based on the single translation hypothesis, two heuristics and AMs are applied in order to improve effectiveness. Firstly, we identify pairs of candidates in which the candidate of each segment is equal to the segment itself (or "one-to-one candidate segments"); for example, headings like *Extracción de terminología / Terminologia-erauzketa* ("terminology extraction"); after that, we remove from the DB the rest of the bigrams which share a same equivalent. Secondly, we identify the cognates whose LCSR > 0.8. For the calculation of LCSR among MW units, we have taken into account the component order; for example, the pair *Inteneteko konexio / conexión a Internet* ("Internet connection") has a low LCSR if both terms are taken as continuous strings. In order to reveal the real cognateness, we calculate the value of LCSR

for the two MW candidates as the sum of the LCRS values for single components of maximal LCSR value. We remove the bigrams as in the previous step. Finally, the remaining bigrams are ranked by one AM like MI, LR, Dice, etc.

3 User interface

The user interface is designed as a corpus navigator that offers the user the possibility of navigating through the results (list of equivalent terms) in their contexts (translation units), and to validate and export the correct equivalences. Different parameters can be chosen to display the results: several AMs, different thresholds (AM, frequency, number of candidates).

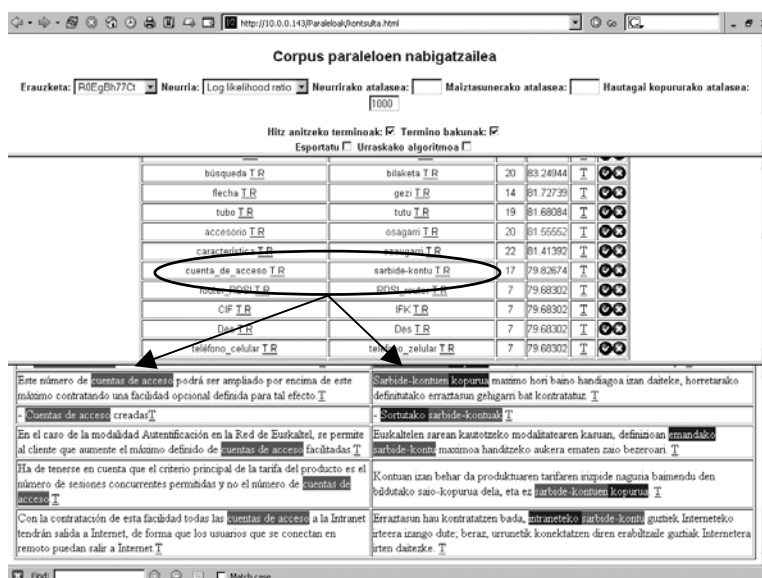


Figure 1. User interface.

The user can validate the correct equivalences and export them to a text file. All the equivalents of a candidate from the list can be displayed (R button). This information is useful to interpret wrong 'best' equivalences that are a result of indirect association. For example, the ones that are related to the fact that some nested candidates in one language can yield higher AM values than maximal NP when term variation is involved.

For example, in the next figure we can see that the nested term *dirección* ("address") is proposed as a equivalent for *helbide elektronikoa* ("electronic address"), but the Spanish terms used in the corpus are *dirección de correo* ("mail address") and *dirección de correo electrónico* ("electronic mail address"). Further refinement of monolingual extraction processes and selection algorithm should provide a better approach to this problem. At the moment, the user can select in the ranking list the correct equivalents for a given term.

Corpus paraleloen nabigatzailea

direccion T R	helbide_elektroniko T R		
		10	54.4697
	hileko_kuota T R	13	54.3263
	lasterbide T R	8	54.3126
	lanqale T R	22	54.1285

direccion	helbide_elektroniko	54.46975
correo_electronico	helbide_elektroniko	25.55137
direccion_de_correo	helbide_elektroniko	20.11892
configuracion	helbide_elektroniko	19.97866
electronico	helbide_elektroniko	10.91831
antivirus_en_su_direccion_de_correo	helbide_elektroniko	10.91831

Si utiliza el programa de correo electronico Outlook Express para que se aplique el servicio antivirus en su <u>direccion</u> de correo debe modificar la configuracion siguiendo los pasos que le indicamos a continuacion:T	Zure <u>helbide elektronikoan</u> antivirus zerbitzua aplikatzeko Outlook Express programa erabiltzen baduzu, aldatu egin behar duzu haren konfigurazioa, honako pauso hauek emanez: T
(Cada linea que visualiza corresponde a la configuracion de una <u>direccion</u> de correo).T	(Ikusten duzun lerro bakoitza <u>helbide elektroniko</u> baten konfiguraziorari dagokio). T
el nombre que tiene antes de @ en su <u>direccion</u> de correo electronico, tambien en minusculas.T	zure <u>helbide elektronikoan</u> @ ikurraren aurretik duzun izena, hori ere letra xehez idatzita. T
el nombre que tiene antes de @ en su <u>direccion</u> de correo electronico, tambien en minusculas pulsando posteriormente en la opcion «Aceptar» primero de esta pantalla y luego de la anterior.T	" zure <u>helbide elektronikoan</u> @ ikurraren aurretik duzun izena, hori ere letra xehez idatzita; ondoren, sakatu ♦ Aceptar ♦ aukera, lehenik pantaila horretakoa eta, ondoren, aurrekoarena. " T
<u>Direccion electronica</u> .T	<u>Helbide elektroniko</u> : T
Para comprobar que todo funciona correctamente, envíe un mensaje a su propia <u>direccion</u> de correo y compruebe que lo recibe correctamente.T	Guztia behar bezala dabilela egiaztatzeko, bidali mezu bat zure <u>helbide elektroniko</u> ra eta egiaztatu behar bezala jaso duzula. T

Figure 2. Ranking of equivalents for a candidate of a given language.

4 Evaluation

A manually tagged translation memory from Euskaltel has been used as a reference for evaluation (10.900 segments; es: 153.163 words; eu: 110.165 words). Figure 3 represents the precision results for the first 5,000 selected pairs of candidates. The best results are obtained using LR (up to 80% precision).

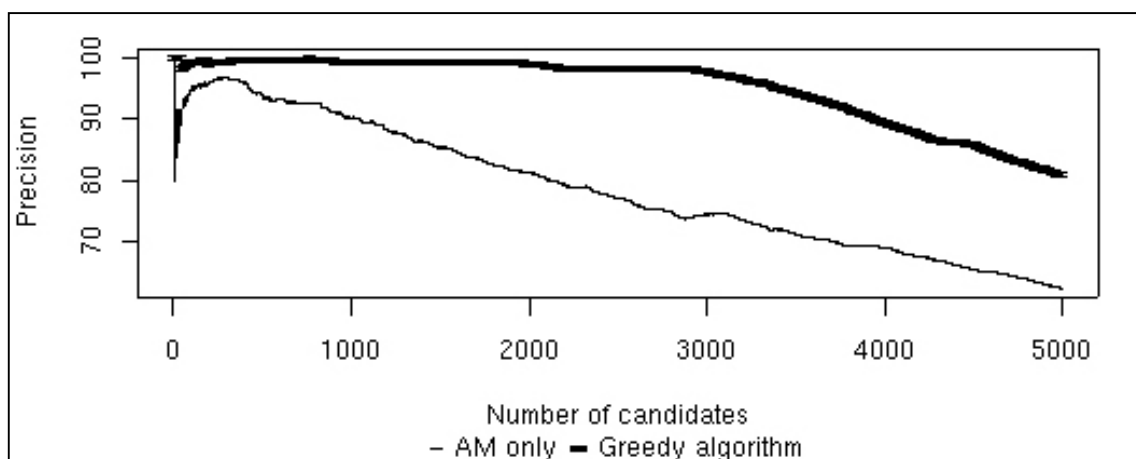


Figure 3. Precision results for the first 5,000 best pairs of candidates.

Precision results are quite good. Candidates pairs from single term segments are very reliable equivalents, and cognates with LCSR > 0.8 show very good performance; candidates with cognates in reversed order are extracted efficiently (see figure 4). Thus, the use of both heuristics in a greedy algorithm based on the single translation hypothesis is adequate.

Source Term	Target Term	Frequency	Other
fiesta I_R	fiesta I_R	6	Kognatuak T
informática I_R	informatika I_R	6	Kognatuak T
condición I_R	kontzizio I_R	6	Kognatuak T
experiencia I_R	esperientzia I_R	5	Kognatuak T
operador de telecomunicación I_R	telekomunikazio-operadore I_R	5	Kognatuak T
especialidad I_R	espezialitate I_R	5	Kognatuak T
gestión I_R	gestio I_R	5	Kognatuak T
transferencia de dato I_R	datu-transferentzia I_R	5	Kognatuak T
sincronización I_R	sinkronizazio I_R	5	Kognatuak T
escritura I_R	eskritura I_R	5	Kognatuak T
acción I_R	akzio I_R	5	Kognatuak T
secreto I_R	sekretu I_R	5	Kognatuak T

Text below the table: "Que Euskaltel, S.A. se ha constituido como operador de telecomunicaciones en el País Vasco..."

Figure 4. Cognate extraction: candidates pairs with cognates in reverse order.

However, the performance with association measures is more irregular than the previous similarity functions, due to the presence of indirect associations. Some of them are natural. Others are caused by the use of different term extractors that produce a little imbalance on the bigram extraction. Despite this problem, the use of monolingual term extractors is satisfactory for the detection of the equivalences between multiword terms or terms with different word length (1:1, n:m and n:m). The next figure shows the case of *número de móvil* / *telefono mugikorraren zenbakia* ("mobile phone number") and *llamada en conferencia* / *konferentzia-dei* ("conference call").

Source Term	Target Term	Frequency	Score
capacidad de 10_Myte T_R	aplikazio/kontu T_R	5	27.83
instalador T_R	instalatzaile T_R	5	27.83
Fundaciones del País Vasco T_R	euskal herriko fundazio T_R	5	27.74
llamada en conferencia T_R	konferentzia-dei T_R	4	27.63
número de móvil T_R	telefono mugikorraren zenbaki T_R	4	27.61
corrección T_R	zuzentasun T_R	3	27.59
Redes T_R	Redes T_R	4	27.42
Mejora Continua T_R	etengabeko hobekuntza T_R	7	27.42

Text below the table: "El fax utilizará una carátula de Amena con su número de móvil. T" and "Faxak Amenaren karatula bat erabiliko du zure telefono mugikorraren zenbakiarekin. T"

Figure 5. Extraction of n:m pairs.

On the other hand, it should be noted that recall is not so satisfactory (50%). This could be partly explained by the fact that the expected bilingual extraction recall is the product of both monolingual recalls. As *Freeling's* recall is 82% and *Erauzterm's* is 85%, the maximum recall that the bilingual extraction could reach is 65%.

5 Conclusions and future work

ELEXBI provides the lexicographer and terminologist with a basic tool for bilingual lexicon extraction from Spanish-Basque translation memories. Precision results are acceptable, but to get better recall it is necessary to refine the linguistic techniques for the detection of candidates both in Basque and Spanish in the monolingual term extractors. In addition to this, the algorithm will be improved to extract more than one equivalent for each term. In this sense, we are now trying to integrate Giza++ (Och, 2002) in our extraction algorithm. In order to improve the precision level we plan to integrate heuristics which use other possible correlated information like morphosyntactic patterns and nesting-level information. Finally, with the aim of being fully operative for lexicographic or terminological work, further integration of the tool is needed (extraction edition, manual extraction, use of pre-existing bilingual lexical resources...).

References

- ALEGRIA I., GURRUTXAGA A., LIZASO P., SARALEGI X., UGARTETXEA S. & URIZAR R. 2004a. "Linguistic and Statistical Approaches to Basque Term Extraction." In *GLAT-2004: The Production Of Specialized Texts*. [on line] [06-03-07] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1079630425/publikoak/Term_Erauzketa.pdf>
- ALEGRIA I., GURRUTXAGA A., LIZASO P., SARALEGI X., UGARTETXEA S. & URIZAR R. 2004b. "A Xml-Based Term Extraction Tool for Basque." In *LREC2004: 4 Th International Conference On Language Resources And Evaluation*. [on line] [06-03-07] http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1078851980/publikoak/Erauzterm_LREC
- CARRERAS, X., CHAO, I., PADRÓ, L., & PADRÓ, M: 2004. "FreeLing: An Open-Source Suite of Language Analyzers" In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.
- FUNG, P. 1998. "A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora". In *Lecture Notes in Artificial Intelligence AMTA 98.*, Springer Publisher, 1998, vol 1529, 1-17. orr.
- KRAIF, O. 2002a. "Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné." *Alignement lexical dans les corpus multilingues, Lexicometrica*, Jean Véronis ed., [on line] [06-03-07] <<http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm>>
- KRAIF, O. 2002b. "Translation alignment and lexical correspondence." In: ALTENBERG, B. & GRANGER, S. (ed.) *Lexis in Contrast*. Amsterdam: John Benjamins.
- MELAMED, I.D. 2001. *Empirical Methods for Exploiting Parallel Texts* MIT Press.
- OCH, F.J & NEY, H. (2002) "Improved Statistical Alignment Models". In *Proc. of the 38th Annual Meeting of the ACL*, pp. 440-447, Hongkong
- SOMERS, H. 2001 *Bilingual Parallel Corpora and Language Engineering*. Language Engineering. *Anglo-Indian Workshop "Language Engineering for South-Asian Languages" (LESAL)*. Mumbai- [on line] [06-03-07] <<http://www.emille.lancs.ac.uk/lesal/somers.pdf>>
- TIEDEMANN, J. 2003. *Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD. Uppsala: Acta Universitatis Upsaliensis. [on line] [06-03-07] <<http://stp.ling.uu.se/~joerg/phd/html/>>