Aligning tagged bitexts

Raquel Martínez
Departamento de Informática y Programación. Facultad de Matemáticas
Universidad Complutense de Madrid, 28040 Madrid, Spain
raquel@eucmos.sim.ucm.es

Joseba Abaitua
Facultad de Filosofía y Letras, Universidad de Deusto
48080 Bilbao, Spain
abaitua@fil.deusto.es

Arantza Casillas
Departamento de Automática, Universidad de Alcalá de Henares
28871 Alcalá de Henares, Spain
arantza@aut.alcala.es

Abstract

This paper describes how complementary techniques can be employed to align multiword expressions in a parallel corpus or bitext. The bitext used for experimentation has two main features: (i) it contains bilingual documents from a dedicated domain of legal and administrative publications rich in specialized jargon; (ii) it involves two languages, Spanish and Basque, which are typologically very distinct (both lexically and morpho-syntactically). The former feature provides a good basis for testing techniques of collocation detection. The latter presents quite a challange to a number of reported algorithms, in particular to the alignment of sentence internal segments.

1 Tagged bitexts as large language resources

Much literature has been produced in the area of sentence alignment of parallel biligual corpora or bitexts. Fewer references concern the alignment of intra-sentential segments such as word or multiword collocations (Eijk 93), (Kupiek 93), (Dagan & Church 94), and (Smadja et al. 96). The difficulty of aligning bitexts depends of a number of factors such as the quality of the bitext (whether is truely parallel or not), the proximity between the languages (either structurally, morpho-syntactically or alphabetically), the additional coded information that bitexts may have (richer or poorer mark-up), among others.

While stuying bitext alignment techniques it was decided that an optimal approach was to tag the corpus. Descriptive annotations can account for linguistic information at all levels, from discourse structure to phonetic features, as well as semantics, syntax and morphology. The process of annotating the corpus in this manner is very labour intensive, even when largely automated, however it produces rewarding results. Thoroughly tagged bitexts become rich and productive language resources (Abaitua et al. 98). SGML based TEI conformant mark-up (Ide & Veronis 95) has been the adopted mark-up option and it was discussed in (Martínez et al. 97).

Continuing with the work of (Martínez et al. 98), where sentence alignment based on rich mark-up was described, we present here two further achievements. Section 2 shows how proper names have been aligned, and Section 3 presents the techniques employed in attempting the aligning of multiword collocations. Results are evaluated in Section 4 and Section 5 offers some discussion.

2 Proper name alignment

2.1 Proper name tagging

The module for the recognition of proper names relies on patterns of typography (capitalisation and punctuation) and on contextual information. It also makes use of lists with most common person, organisation, law, publication and place names. The tagger annotates a multiword chain as a proper name <rs> when each word in the chain is uppercase initial. A closed list of functional words (prepositions, conjunctions, determiners, etc.) is allowed to appear

inside the proper name chain, see examples in Table 1. A collection of heuristics discard uppercase initial words in sentence initial position or in other exceptional cases.

Just us (Smadja et al. 96) distinguished between two types of collocations, we too distinguish between:

- **Fixed names**: Compound proper names labeled 'fixed', such as *Boletín Oficial de Bizkaia*, are rigid compounds. Spanish proper names all correspond to this type.
- **Flexible names**: Compound proper names labeled 'flexible' are compounds that can be separated by intervening text elements such as in *Administrazio Publikoetarako Ministeritzaren* <date>.. </date> *Agindua*, where a date splits the tokens within the compound. There is a small but significative number of these in Basque, as has been previously noted by (Aduriz et al. 96b).

After proper names have been successfully identified (Table 2), the next step is their alignment. Two types of alignment can take place:

- **1 to 1 alignment**: one to one correspondence between fixed names in the source and target documents.
- **1 to N alignment**: one to none or more than one correspondences between fixed names in the source language and flexible names in the target language.

Alignment has been achieved by resorting to:

1. Proper name categorization, as shown in Table 1.
2. Reduction of the alignment space to previously aligned sentences.
3. Identification of cognate nouns, aided by a set of phonological rules that apply when Basque loan terms are directly derived from Spanish terms.
4. The application of the *TasC* algorithm (Martínez et al. 98) adapted to proper name alignment.

2.2 Identification of cognates

Points one and two above may suffice to work up the alignment of fixed proper names belonging to a single category that shows up only once in the alignment space (i.e. in the sentence). Nevertheless, there can be sentences with flexible proper names or more than one fixed proper name belonging to the same category. Therefore it may be necessary to determine the correct alignment among possible candidates. As additional criteria in these cases we reinforce the identification of lexical cognates with a set of phonological correlation rules.

These are two examples of phonological correlation rules:

(i) The Spanish prefix '*rel-*' always correlates with the prefix '*erl-*' in Basque loans (e.g. *reloj / erloju; relación / erlazio*)

(ii) The Spanish suffix '*-ción*' often correlates with the suffix '*-zio*' in Basque loans (e.g. *noción / nozio; administración / administrazio*)

We use a set of up to 33 rules of this type. For some loan terms in Basque e.g. *universidad / unibertsitate*), several of these rules may apply: *-v-* → *-b-* ; *-rs-* → *-rts-* ; *-dad* → *-tate.*

Although the application of these phonological rules for identifying Basque loan words is quite regular, not every new term in Basque is derived in this way. In many other cases a Spanish term has a genuine Basque counterpart, (e.g. *sociedad / elkarte*). In any case, this set of phonological rules provides a very efficient aid for the identification of a high proportion of Spanish/Basque cognates (86.45 % on average, as shown in Table 3).

Therefore, when aligning proper names, cognate identification will help not only in obvious cases such as personal or place names, but also in categories of proper names such as organization, law or title.

2.3 Calculating the similarity between Basque and Spanish proper names

In order to determine whether two proper names belonging to the same category are translation equivalencies of each other, Dice's coefficient (Dice 45) is applied in two phases: first at token level and then, at proper name level.

Categories	Spanish	Basque
Person	*Javier Otazua Barrena*	*Javier Otazua Barrena*
Place	*Amorebieta-Etxano* *Corredor del Cadagua* *Bilbao c/ Alameda Rekalde*	*Amorebieta-Etxano* *Kadaguako pasabidea* *Bilboko Errekalde Zumarkaleko*
Organization	*Ayuntamiento de Areatza* *Registro de la Propiedad* *Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco*	*Areatzako udalak* *Jabegoaren erroldaritzan* *Euskal Herriko Justizia Auzitegi Nagusiko Administraziozko Liskarrauzietarako Salari*
Law	*Impuesto sobre la Renta* *Plan Especial de Reforma Interior* *Normativa de Rehabilitación*	*Errentaren gaineko Zergari* *Barne-Eraberritzearen Plan Beretzia* *Birgaikuntzari buruzko Arauko*
Title	*Jefe del Servicio de Administración de Tributos Directos* *Diputado Foral de Hacienda y Finanzas*	*Zuzeneko Zergen Administrazio Zerbitzuko buruaren* *Ogasun eta Finantzen foru diputatua*
Publication	*Boletín de Bizkaia* *Boletín Oficial del País Pasco*	*Bizkaiko Aldizkari* *Euskal Herriko Aldizkari Ofizialean*
Uncategorized	*Estudio de Detalle* *Acción Comunitaria* *Documento Nacional de Identidad*	*Azterlan Zehatzarako* *Erkidego Ekintzapidearen* *Nortasun Agiri Nazionalaren*

Table 1: Examples of proper names

	Spanish			Basque		
Proper Name Classes	Precision	Recall	% Spanish PN	Precision	Recall	% Basque PN
Person	100%	100%	4.48%	100%	100%	4.76%
Place	100%	100%	6.38%	100%	100%	6.95%
Organisation	99.2%	97.8%	23.96%	100%	100%	24.17%
Law	99.2%	99.2%	47.93%	100%	100%	46.15%
Title	100%	100%	6.55%	97.2%	97.2%	6.59%
Publication	100%	100%	2.58%	100%	100%	2.74%
Uncategorised	100%	100%	8.10%	100%	100%	8.60
Total	99.4%	99.1%	100%	99.8%	99.8%	100%

Table 2: Results of proper name identification

1. In the first level, each token in the source proper name is compared with all the tokens in the target proper name. In order to determine whether two tokens are cognates, bigrams are compared trying to apply, if they are not equal, the rules of phonological derivation. Only when the resulting coefficient is bigger than a threshold, the tokens are considered cognates. The threshold has been established in 0.5 as a result of different experimental tests.

2. In the second level, given a source and a target proper name, their similarity is determined according to the number of cognate tokens that exist between them. Figure 2 we illustrates an instance of the application of Dice's coefficient at both levels.

Spanish proper name ——— Basque proper name
Boletín Oficial de Bizkaia — Bizkaiko Aldizkari Ofizialean

First level of similatity:
Boletín —— none
Oficial —— *Ofizialean* (*-c-* → *-z-*)
DC= $\frac{2\times6}{6+9}$ = 0.8 > 0.5
Bizkaia —— *Bizkaiko* (no rule)
DC= $\frac{2\times5}{6+7}$ = 0.76 > 0.5

Second level of similatity:
Number of cognate tokens is 2, then:
DC = $\frac{2\times2}{3+3}$ = 0.66

Figure 2: Example of similarity calculation between two proper name

2.4 Algorithm for proper name alignment

After similarity metrics have been set between candidate proper names, the alignment algo-

<s id=sESdoc12-2 > Segundo: Notificar la presente <rs type=law> *Orden Foral* </rs> al <rs type=organisation > *Ayuntamiento de Barrika* </rs>, publicarla en el <rs type=publication >*Boletín Oficial de Bizkaia* </rs> y proceder a la autentificación del <rs id= type=law > *Plan Parcial* </seg> tal como ha sido presentado. </s>

<s id=sEUdoc12-3 > Bigarrena: Honako erabakia <rs type=organisation > *Barrikako Udalari* </rs> jakinerazi eta <rs type=publication > *Bizkaiko Aldizkari Ofizialean* </rs> argitaratzea eta <rs type=law >*Plan Partziala* </rs> aurkeztua izan den eran kautotzea. </s>

Figure 1: Example of non-literal translation

rithm can be applied. The alignment algorithm has been borrowed from (Martínez et al. 98) with minor differences.

- The first difference is the criteria by means of which the similarity amongst alignable candidates is determined. While sentences are aligned on the basis of the similarity of the annotations they contain, proper names are aligned on the basis of their belonging to the same category as well as by matching cognate tokens.
- The second difference is the relevance of the order of alignable elements in the bitext. While in sentence alignment there are constraints regarding ordering and grouping to reduce the number of cases to be evaluated, in the aligning of proper names constraints cannot be applied because ordering is not predictable.

Due to non-literal translations, 12% of the identified proper names have no exact counterpart in the other language (see Figure 1). In this case, the Basque sentence does not have the proper name of the law *Orden Foral* but the anaphoric nominal *honako erabakia*, 'this resolution'. In the corpus, there are 6% more proper name in the Spanish side of the bitext than in the Basque side.

Table 3 shows the accuracy of this alignment strategy. Proper names with no counterpart have not been considered. Figure 3 illustrates an instance of how aligned proper names are tagged in the bitext.

3 Alignment of collocations

(Smadja et al. 96), (Dagan & Church 94), (Kupiek 93) and (Eijk 93) approach the alignment of multiword collocations resorting to a number of complementary techniques:

(i) **Noun phrase collocations**: All but Smadja narrow the scope of collocations to noun phrases. Smadja is the only one that attempts to treat other phrases (such as verb phrases as well what he labels 'flexible phrases').

(ii) **Delimited search space**: All but Church delimit the search space to already aligned sentences. Church in turn departs from a corpus of aligned words.

(iii) **POS tagging**: All but Smadja employ Part of Speech (POS) taggers.

We also employ techniques (ii) and (iii), but we introduce three additional resources: A bilingual glossary, a bilingual contrastive grammar, as well as the structural markup which alredy exists in the bitext. In addition, we also consider verb prases.

The approach discribed below illustrates work in progress on how we try to optimize the alignment process by combining those techniques. Collocations are aligned in six steps. The first three steps are meant to detect candidate collocations in both languages. The last three are directly involved in the alignment.

1. **Word cooccurrence frequency**: Due to the specialized nature of the bitext, any word cooccurrence that superates a given threshold is considered to be a collocation candidate. This threshold depends on the size of the corpus, but even a low figure as 2 can be considered significative enough. A tool for word cooccurrence detection has been implemented. This tool is sensitive to SGML tags and it uses a window of maximum ten words. From a subcorpus of 150,000 words, with a threshold of 3 and a windows size of 7, 2,095 candidate colloca-

Categories	% Alignable PN	Precision	Recall
Person	100%	100%	100%
Place	89.28%	100%	92%
Organisation	79.38%	96.7%	76.6%
Law	95.68%	100%	88.2%
Title	86.2%	100%	72.3%
Publication	100%	100%	100%
Uncategorised	54.54%	93.4%	85.7%
Total	86.45%	98.5%	87.82%

Table 3: Results of proper name alignment

Spanish Sentences:
<s id=sES734 corresp=sEU740> <num num=1> 1. </num> Suspender la aprobación definitiva de la <rs type=law id=LES367 corresp=UEU141,LEU342> ***Modificación Puntual de las Normas Subsidiarias de Planeamiento Municipal de Gernika-Lumo en el Barrio de Arana*** </rs>, en base a las deficiencias que a continuación se expresan y que debeán subsanarse <colon> : </colon> </s>

Basque Sentences:
<s id=sEU740 corresp=sES734> <num num=1> 1. </num> <rs type=place id=UEU141 corresp=LES367> ***Araneko Auzoan*** </rs>, <rs type=law id=LEU342 corresp=LES367> ***Gernika-Lumoko Udal Egitamuketazko Ordezko Arauen Puntuzko Aldaketaren*** </rs> behin betirako onarpena etetzea, jarraian adierazten diren eta zuzendu egin beharko diren akatsetan oinarrituta <colon> : </colon> </s>

Figure 3: Sample of 1 to N alignment of proper names

tions in Spanish and 1,483 in Basque have been detected .

2. **POS tagging**: A tagged version of the Spanish text was supplied by the Natural Language Research Group at the *Universitat Politècnica de Catalunya* (Márquez & Padró 97). The Basque text was tagged by the IXA group from the *Euskal Herriko Unibertsitatea* (Aduriz et al. 96a), (see Figure 5).

3. **NP, VP grammars**: Simple noun phrase and verb phrase patterns have been used to detect candidate collocations and to filter out inapproppriate word cooccurrences. By means of this technique, 80% of the detected word cooccurrences are discarded. Basque and Spanish phrases show great divergences, and for the alignment procedure to succed, it has been necessary to implement an additional resource: a correspondence table with grammatical patterns for Spanish and Basque phrases (see Table 4).

4. **Bilingual glossary lookup**: This is a very useful resource containing over 15,000 aligned entries. The glossary was developed by the same translators that were in charge of the corpus we are working with. Yet, the glossary, although it is available on-line, translators have not applied it systematically and frequent divergences arise (compare Figure 4 with Figure 6).

5. **Search within aligned sentences**: Aligned senteces delimit the search space thereby reducing the complexity of the alignment.

6. **Human validation**: The final step involves human intervention, so that detected collocations can be validated and thus incorporated into the glossary. The posibility of enriching the glossary with contextual information has not yet been implemented, but holds great potentiallity (<doctype>, , <p> and <s> tags could be used to locate collocations in context and index them through their correspondig id tag attributes).

4 Evaluation

Scores of proper name alignment are shown in Table 3 and are very satisfactory. With regards to collocations, we expected that those candidate collocations found in the bilingual glossary would show high alignment scores, which has

Spanish	Basque
N+	N+
N A+	N A+
N A+	N N+
N P N	N+
(aux) V+	V+ (aux)
etc.	

Table 4: Correspondence table

been the case. We still do not have definite estimations on the performance of collocations not present in the glossary. As we discuss below, we are still sceptic about the results of the correspondence table with current version of the Basque lemmatizer.

5 Discussion

We have not yet calculated how many detected collocations are included in the glossary, although it has become clear that a high proportion of these detected collocations have not been considered by the translators who created the dictionary. These tend to include only collocations which have a clear terminological appearance. It is hard to discriminate between general language collocations and domain specific terminology and this discussion is beyond the scope of this paper.

The correspondence table with Spanish and Basque grammatical patterns is at present problematic. This is due to the lack of morphological information in the output of the Basque lemmatizer. Basque is an aglutinative language which has postpositions and other functional elements added as suffixes. The information such suffixes provide is not shown by the lemmatizer and this inevitably hinders the efficiency of the correspondence table. However we are confident that future versions of both the Basque and Spanish lemmatizers will become closer because they are currently developed within the same project team. When their output becomes more homogeneous, the efficiency of the correspondence table will be greately increased.

6 Acknowledgements

This research is being partially supported by the Spanish Research Agency, project ITEM, TIC-96-1243-C03-01. We greately appreciate the help given to us by Felisa Verdejo, director of the project. We are particulary endebeted

```
agotar ; agortu
"agotar el plazo";"epea agortu"
"agotar la va administrativa";
"administrazio bidea agortu"
"agotarse las reservas";"erreserbak agortu"
...
"defender";"defendatu"
"defender los derechos";"eskubideak defendatu"
"defensa";"defentsa"
"defensa civil";"defentsa zibil"
...
"interponer";"jarri"
"interponer un recurso";"errekurtsoa jarri"
"interponer una reclamación";"erreklamazioa jarri"
"interpretación";"interpretazio"
...
"medidor";"neurgailu"
"medio";"1) bide; 2) eskuarte"
"medio audiovisual";"ikusentzunezko helbide"
"medio de comunicacin";"komunikabide"
...
"recurso administrativo";"administrazio errekurtso"
"recurso contencioso administrativo";
"Administrazioarekiko auzibide-errekurtso"
"recurso de abuso";"abusu errekurtso"
...
"veto";"geben"
"vía administrativa";"administrazio bide"
"vía administrativa, por";"administrazio bidetik"
...
```

Figure 6: Glossary sample

to developers of the Spanish (Márquez & Padró 97) and the Basque (Aduriz et al. 96a) lemmatizers. We thank CRL for allowing us the use of their premises and to Begoña Farwell for the reviewing of the text.

References

(Abaitua et al. 97) J. Abaitua, A. Casillas, R. Martínez. Value Added Tagging for Multilingual Resource Management. *Proceedings of the First International Conference on Language Resources & Evaluation*, ELRA, 1003-1007, 1998.

(Aduriz et al. 96a) I. Aduriz, I. Aldezabal, I. Alegría, R. Urizar. EUSLEM: A lemmatiser/tagger for Basque. *EURALEX'96*, Gotteborg, Sweden, 1996.

(Aduriz et al. 96b) I. Aduriz, I. Aldezabal, X. Artola, N. Ezeiza, and R. Urizar. MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque *Papers in Computational Lexicography COMPLEX'96*, 1-8. Budapest 1996.

(Dagan & Church 94) I. Dagan, K. W. Church. Termight: Identifying and Translating Technical Terminology. *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLP-94, 34-40, Stuttgart, Germany, 1994.

 <s> Contra dicha <rs type=law id=LES546 corresp=LEU540> *Orden Foral* </rs>, que agota la <term id=X1 corresp=X1> **vía administrativa** </term> <term id=X2 corresp=X3> **podrá interponerse** </term> <term id=X3 corresp=X2>***recurso contencioso-administrativo*** </term> ante la <rs type=organization id=0ES867 corresp=0EU856> *Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco* </rs>, en el plazo de dos meses, contado desde el día siguiente a esta notificación, sin perjuicio de la utilización de otros <term id=X4 corresp=X4>***medios de defensa*** </term> que estime oportunos. </s>

 <s> <rs type=law id=LEU540 corresp=LES546> *Foru Agindu* </rs> horrek amaiera eman dio <term id=X1 corresp=X1> ***administrazio bideari*** </term>; eta beraren aurka <term id=X2 corresp=X3> ***administrazioarekiko auzibide-errekurtsoa*** </term> <term id=X3 corresp=X2>**jarri ahal izango zaio** </term> <rs type=law id=0EU856 corresp=0ES867> *Euskal Herriko Justizi Auzitegi Nagusiko Administrazioarekiko Auzibideetarako Salari* </rs>, bi hilabeteko epean; jakinarazpen hau egiten den egunaren biharamunetik zenbatuko da epe hori; hala eta guztiz ere, egokiesten diren beste <term id=X4 corresp=X4>***defentsabideak*** </term> ere erabil litezke. </s>

Figure 4: Example of aligned collocations

Spanish lemmatization output:

...
que que B3323 PR3CN000
agota agotar 6202030 VMIP3S0
la la 810 TDFS0
vía vía 010 NCFS000
administrativa administrativo 110 AQ0FS00
podrá poder 6202330 VMIF3S0
interponerse interponer 6223503 VMN0000
recurso recurso 000 NCMS000
contencioso-administrativo
contencioso-administrativo NOMASK AQ00000
ante ante A1 SPS00
la la 810 TDFS0
...
sin-perjuicio-de sin-perjuicio-de A1 SPS00
la la 810 TDFS0
utilización utilización 010 NCFS000
de de A1 SPS00
otros otro 3012 DI3MP00
medios medio 001 NCMP000
de de A1 SPS00
defensa defensa 010 NCFS000
que que B3323 PR3CN000
estime estimar 6202032 VMMP3S0
oportunos oportuno 101 AQ0MP00
...

Basque lemmatization output:

...
amaiera amaiera IZE-ARR
eman eman ADI-SIN
dio *edun ADL
administrazio administrazio IZE-ARR
bideari bide IZE-ARR
; EOS = PUNT-PUNTU
eta BOS eta LOT-JNT
beraren bera IOR-PER
aurka aurka IZE-ARR
administrazioarekiko administrazio IZE-ARR
auzibide-errekurtsoa auzibide-errekurtso IZE-ARR
jarri jar ADI-SIN
ahal ahal ADI-ADP
izango izan ADI-SIN
zaio izan ADL
...
hala BOS hala ADB-ADO
eta eta LOT-JNT
guztiz guztiz MAI
ere ere LOT-LOK
, = PUNT-EZPUN
egokiesten (egokiesta) IZE-ARR
diren izan ADT
beste beste DET-DZG
defentsabideak defentsabide IZE-ARR
ere ere LOT-LOK
erabil erabil ADI-SIN
litezke *edin ADL
. EOS = PUNT-PUN

Figure 5: Output of both Spanish and Basque lemmatizations

(Dice 45) L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26, 297-302.

(Eijk 93) P. van der Eijk. Automating the Acquisition of Bilingual Terminology. *Proceedings Sixth Conference of the European Chapter of the Association for Computational Linguistic*, Utrecht, The Netherlands, 113-119, 1993.

(Ide & Veronis 95) N. Ide, J. Veronis. *The Text Encoding Initiative: Background and Contexts*. Dordrecht: Kluwer Academic Publishers, 1995.

(Kupiec 93) J. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. *Proceed-*

bus, Ohio, 17-22. Association for Computational Linguistics 1993.

(Márquez & Padró, 97) L. Márquez, L. Padró. A Flexible POS Tagger Using an Automatically Acquired Language Model. *Proceedings of the joint EACL/ACL97*, Madrid, Spain, 1997.

(Martínez et al. 97) R. Martínez, A. Casillas and J. Abaitua. Bilingual parallel text segmentation and tagging for specialized documentation. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP'97, 369-372, 1997.

(Martínez et al. 98) R. Martínez, A. Casillas and J. Abaitua. Bitext Correspondences through Rich Markup. *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98) and 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, Montreal, Canada, 1998.

(Smadja et al. 96) F. Smadja, K. McKeown, V. Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics* Volume 22, No. 1, 1996.