Value Added Tagging for Multilingual Resource Management

Joseba Abaitua
Facultad de Filosofía y Letras
Universidad de Deusto, 48080 Bilbao, Spain
`e-mail:abaitua@fil.deusto.es`

Arantza Casillas
Departamento de Automática. Universidad de Alcalá de Henares
28871 Alcalá de Henares, Spain
`e-mail:arantza@aut.alcala.es`

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

Abstract

The Legebiduna project brings together state-of-the-art techniques in multilingual corpus management, generic mark-up, text segmentation and alignment, terminological extraction, automatic text cataloguing, and reutilisation of recurrent text in specialised documentation. We report on the experience of a four year project of bilingual corpus mining in a dedicated domain of official bilingual publications. Considerable effort has been made in developing tools for the automatic processing of a collected parallel corpus of 7 million words in both Spanish and Basque. Experiments have been undertaken on a half million word sample of the corpus, and the results are very satisfactory. Legebiduna has now become a prototype of a domain-expert editing tool that helps both institutional writers and translators to carry out their work in an optimal computer oriented environment.

1 Introduction

Producing bilingual documentation within specialised domains is a very time-consuming and expensive process. It is furthermore a relatively unautomated task, in spite of its potentialities. In the manual process it involves both human writers and translators, who devote endless efforts in a constant recycling of repetitive and reusable text chunks. The main desire of institutional writers as well as translators is to quickly ascertain how a recurrent text (whether memo, resolution, announcement, etc.) has been previously composed so as to save the effort of attempting a novel and possibly problematic unseen version. Textual variations and divergences are not much appreciated in specialised documentation. When there is evidence than a similar document might have been previously written or translated, they take pains to find it in the normalised version. The manual process spans through several difficult and inefficient steps: First, the source document type (resolution, announcement, etc.) is analysed and recognised. Then, a large and usually not very well organised set of folders, spread out in different disk units, are visually scanned with the hope that a similar document-token can be found. If this succeeds, the document is retrieved, and a battery of editing operations (cut, copy and paste) take place so that reusable fragments are refitted together with new information. Finally, obsolete data, such as dates, proper nomenclature, or numbering is updated.

Advances in computational linguistics, machine translation, electronic publishing and data mining can be put together at their best performance to clone such repetitious and costly processes within an appropriate computer-oriented environment. The Legebiduna project combines the creation of translation memories from a bilingual automatically tagged corpus with a SGML-based editing tool for source-document generation. There are similar reported experiments in the literature, but none of them fully integrates both processes of writing and translating the same document in a single system.

2 Parallel Bilingual Corpus

The scope of application of Legebiduna has been restricted to the domain of institutional publications issued by the Basque local administration in Spain. Since the declaration of official bilingualism in 1980, institutional bulletins in the Basque region must be published both in Spanish and Basque. Around 300 human translators are devoted to the hard task of translating over 70,000 pages per year of official publications. This in itself represents a high proportion (over 80%) of the demand for translating into Basque. Yet, institutional sources have reported that no more than a 20% of the total administrative documentation reaches the translation stage.

A bilingual corpus of over 7 million words in each Spanish and Basque has been collected. However, due to severe noise problems (missing fragments, awkward formats, mark-up miscellany, etc.) it has not been possible to work with the whole collection, and for the sake of prototyping we have selected a representative subset of around 500,000 words of parallel texts in pure ASCII format without any usable annotation.

All document instances have been automatically marked up and accurately processed on the basis of multistrata cycles, ranging from more general mark-up (paragraph, sentences, quoted text), through document specific tagging (text headers, divisions, identification codes), up to more linguistically oriented mark-up (terms, proper names, collocations). Mark-up also stands for the alignment of parallel text segments.

Corpora containing bilingual versions of the same text entity have been called "bitexts". Annotated bitexts are a very useful source of data for applications such as example and memory based machine translation (Sumita & Iida, 1991; Brown et al., 1993; Collins et al., 1996); bilingual terminology extraction (Kupiec, 1993; Eijk, 1993; Dagan et al., 1994; Smajda et al., 1996); bilingual lexicography (Catizione et al., 1993; Daille et al., 1994; Gale & Church, 1991); multilingual information retrieval (SIGIR, 1996; Yang et al., 1997); and word-sense disambiguation (Gale et al., 1992; Chan & Chen, 1997).

Parallel texts in annotated form are becoming increasingly available (e.g. WWW pages of multilingual institutions such as the European Union, United Nations, UNESCO, etc.). Although the mark-up is normally insufficient, it is possible to enrich existing annotations through various tagging phases.

3 Tagging and Segmentation into Translation Units

We have tried to make our approach to bitext processing optimal by the utilisation of a very precise and well-tuned segmentation procedure that recognises translation units (Abaitua et. al., 1997). This consists of a set of subtools that perform such processes as: sentences boundary detection, proper noun tagging, recognition of other text entities such as numbers, dates, abbreviations, enumerations, as well as other document internal logic entities. These subtools can be used independently at various stages of automatic tagging. Based on pattern matching and heuristics, these tools produce different descriptive levels:

- General encoding (paragraph, sentence, quoted text, dates, numbers, abbreviations, etc.), much like the Mtseg tool of MULTEXT (MtSeg, 1997).
- Document specific tags that identify document types and define document internal logic entities (sections, divisions, identification code, number and date of issue, issuer, lists, itemised sections, etc.). A typological study of the corpus was carried out in order to determine the logical structure for each document token in our sample. Pattern matching techniques and heuristics have been used as a way of capturing the internal composition of documents in terms of SGML tags.
- Proper noun tagging. Proper nouns are identified and classified as person, place, organisation, law, title, publication or uncategorised.

Some of this collection of tags (shown in Table 1) reflect basic structural and referential elements, which appear consistently on both sides of the bitext. The encoding scheme has been based on TEI's guidelines for SGML based mark-up (Ide & Veronis, 1995) and has been described in (Martínez et al., 1997). The results of the identification of the description levels are shown in Table 2.

Following (Abaitua et al., 1997), segmentation into translation units is based on the following classification:

1. Formulaic translation units. These typical multi-clause constructions are very frequent in legal and administrative sublanguages. Recognition is carried out by means of straightforward pattern matching techniques.
2. Terminological translation units. These belong to three subgroups:
    - Specialised terminology
      We departed form a specialised bilingual glossary of 15,000 terms compiled by human translators of the Basque Administration. Terms in the glossary have been matched against the corpus and additional items have been included in the glossary. Recognised strings in the corpus have been annotated with the `<term id=X corresp=Y>` tag.
    - Domain specific collocations
      These are recurrent word combinations in the corpus, which at times contain undetected terminology, and occasionally resemble phrasal expressions typical of the domain. Co-occurring items were later filtered out in consecutive steps. First the algorithm of (Frantzi & Ananiadou, 1996) was applied to detect spurious repetitions and nested embedding. Then the results were screened by a stop list (made of prepositions, conjunctions and determiners). Finally, the candidate expressions were POS tagged and matched up against a mini noun phrase grammar. The succeeding noun phrases have been revised by a human terminologist and added up to the specialised glossary and marked-up as `<term id=X corresp=Y>` in the corpus.
    - Proper terms
      These are multiword compounds that correspond to proper names of people, institutions, laws, places, etc. Proper terms have

Descriptive levels	Tagset
General encoding	<p>, <s>, <num>, <date> <abbr>, <q>
Document especific	<div>, <classCode> <keywords>, <dateline>, <list><seg>
Proper nouns	<rs>

Table 1: Tagset used for sentence alignment

	Spanish		Basque	
Descriptive levels	Precision	Recall	Precision	Recall
General	99.7%	99.4%	98.6%	98.5%
Document specific	96.8%	96.3%	95.8%	95.6%
Proper noun	94.4%	99.1%	98.8%	99.8%

Table 2: Results of description levels encoding

Cases	%Corpus	% Accuracy
1 - 1	94.39%	100%
N - M	5.61%	99.68%

Table 3: Sentence alignment algorithm results

been annotated and aligned by means of the `<rs type=X id=Y corresp=Z>` tag.

3. Lexicological translation units. Other generic vocabulary that cannot be recognised as belonging to the specialised domain has not been treated or annotated.

4 Alignment

An algorithm that is not disrupted by word order differences, nor small asymmetries in the bitext has been developed. Unlike other reported algorithms, it possesses the additional advantage of being portable to any pair of languages without the need to resort to any language-specific heuristics. If bitext mark-up is adequate and consistent, sentence alignment becomes a simple and accurate process. One of the best consequences of this approach is that the burden of language dependent processing is carried out during the monolingual tagging and segmentation phases.

The result of sentence alignment is reflected in the bitext by the incorporation of the attribute '`corresp`' to sentence tags, as can be seen in Figure 1. This attribute points to the corresponding sentence identification code in the other language.

The current version of the algorithm has been tested against a subcorpus of 500,000 words in each language consisting of 5,988 sentences and has rendered the results shown in Table 3.

The alignment algorithm has been designed in such a modular way that it can easily change the tagset used for alignment and the weight of each tag to adapt it to different bitext annotations. The current version of the algorithm uses the tagset shown in Table 1 without weights.

4.1 Proper Noun Alignment

Proper nouns in the bitext are aligned within the context of aligned sentences. It is important that proper nouns are adequately aligned: indexation of the translation memory and retrieval of relevant document pieces by the editing tool crucially relies on them. The recognition algorithm distinguishes between two classes of proper nouns:

- Rigid proper nouns. These are rigid compounds such as *Boletín Oficial de Bizkaia*. All the Spanish proper nouns correspond to this category.
- Flexible proper nouns. These are proper nouns that can be separated by intervening text elements such as *Administrazio Publikoetarako Ministeritzaren* `<date>`... `</date>` *Agindua*, where a date splits the tokens of the noun. As has been noticed before (Aduriz et al., 1996), there is a number of Basque multiword expressions that fall under this class.

In non-literal translations, 12% of Spanish proper nouns have no exact counterpart in Basque, yet the output of the alignment process is very successful, as can be seen in Table 4.

4.2 Extending the Alignment Algorithm

We are trying to improve the accuracy rates of proper noun alignment, and the next step is the alignment of collocations. Due to the still unstable translation choices of much administrative terminology in Basque, on top of the considerable typological and structural differences between Basque and Spanish, many of the techniques reported in the literature (Smadja et al., 1996; Kupiec, 1993; Eijk, 1993) cannot be effectively applied. POS tagging combined with recurrent bilingual glossary lookup is the approach we are currently experimenting with.

5 DTD Abstraction

SGML mark-up provides a way to determine the logical structure of a document and its syntax in the form of a context-free grammar. This is called the Document Type Definition (DTD) and it contains specifications for:

Spanish Sentence:
`<s id=sESdoc5-4 corresp=sEUdoc5-5>`Habiéndose detectado en el anuncio publicado en el número`<num num=79>` 79 `</num>` de fecha `<date date=27/04>`27 de abril`</date>` de este `<rs type=publication>`Boletín`</rs>`, la omisión del primer párrafo de la `<rs type=law>`Orden Foral`</rs>` de referencia se procede a su íntegra publicación.`</s>`

Basque Sentence:
`<s id=sEUdoc5-5 corresp=sESdoc5-4>`Agerkaria honetako `<date date=27/04>` apirilaren 27ko`</date>` `<num num=79>`79k.an `</num>` argitaratutako iragarkian aipameneko `<rs type=law>`Foru Aginduaren`</rs>` lehen lerroaldea ez dela geri detektatu ondoren beraren argitarapen osoa egitera jo da.`</s>`

Figure 1: Results of sentence alignment expressed by the `corresp` attribute

Proper Noun Classes	% Alignables 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 4: Results of the alignment of proper nouns

- Names and content for all elements that are permitted to appear in a document.
- Order in which these elements must appear.
- Tag attributes with default values for those elements.

Because the documentation in our corpus was not produced using SGML based editing software, and hence does not comply with any DTD, DTDs have been abstracted away from the annotations that were automatically introduced in the corpus. Similar experiments have been reported before in the literature. (Ahonen, 1995) uses a method to build document instances from tagged texts that consists of a deterministic finite automaton for each context model. Subsequently, these automata are generalised and converted into regular expressions which are easily transcribed into SGML content models. (Shafer, 1995) combines document instances with simplification rules. Our method is similar to Shafer's, but with a modification in the way rules reduce document instances. A tool to obtain a DTD for all document instances has been developed.

In the domain of official documentation, one of the most desired properties is consistency, that is, all different instances of one single document-type must share the same logical structure. The attainment of this property is one of the best spin-offs of the formal constraining force that an SGML's DTD imposes on new documents. Our aim is to provide writers and translators of official documentation with an authoring environment that takes advantage of this property, that is, an editing tool in which the process of generating new bilingual documents is directed by paired DTDs.

6 Translation Memory

Aligned bilingual text segments and DTDs are stored on two databases, one for each of the collections of translated segments identified and aligned in each language, which are indexed by tag names and attributes. Paired DTDs together with the collection of aligned bitext segments constitute the translation memory. This helps both the institutional writer as well as the translator in generating the bilingual document by suggesting the document structure and proposing some logical element contents and translations.

DTDs cannot indicate directly the linguistic content of the elements concurring in a document, but this content can be indirectly linked through an intermediate database, which, as in our case, stores all possible contents for each element in a document.

Text produced via a DTD-based generation grammar inherits its DTD's hierarchical structure and can hence be represented by a graph whose nodes are either elements or other DTDs. The generation process is directed by this graph representation.

7 System Architecture

In a structured editing system, a document is considered as a logical structure. It is made up of typed components such as title, abstract, sections, etc. which are assembled into a structure representing the organisation of the document. The types of components and their relationship in the structure are defined by a generic structure, and each document has a specific structure which is an instance of the generic structure. Several generic structures may be defined to represent different types of documents. This implies that each document must have a specific logical structure which is consistent with the corresponding generic structure.

In the common case, the generation of a SGML doc-

```
<!ELEMENTbody − −(div1, div2, div3) >
<!ELEMENTdiv1 − −(category, ident) >
<!ELEMENTcategory − −(#PCDATA) >
<!ELEMENTident − −(classCode, date) >
<!ELEMENTclassCode − −(#PCDATA) >
<!ELEMENTdate − −(#PCDATA) >
<!ELEMENTdiv2 − −(#PCDATA|seg|abbr|num|seg 9
|seg 10)+ >
<!ELEMENTseg − −(#PCDATA) >
<!ELEMENTabbr − −(#PCDATA) >
<!ELEMENTnum − −(#PCDATA) >
<!ELEMENTseg 9 − −(#PCDATA|seg|abbr|num) >
<!ELEMENTseg 10−−(#PCDATA|seg|abbr|num)+ >
<!ELEMENTseg − −(#PCDATA) >
<!ELEMENTabbr − −(#PCDATA) >
<!ELEMENTnum − −(#PCDATA) >
<!ELEMENTdiv3 − −(docAuthor, dateline?) >
<!ELEMENTdocAuthor − −(#PCDATA) >
<!ELEMENTdateline − −(#PCDATA) >
```

Figure 2: DTD of a document type

ument can be seen as a top-down procedure. Departing from an elected DTD (see Figure 2), which represents the general case of the logical structure for the required document type, a concrete instance of that particular case may be produced.

The editing environment directs the generation of both the source text in Spanish and the target document in Basque through a planification process of the logical order of the document elements and their content. Two levels of text generation may be considered. There is a strategic level of decision which permits to organise the logical structure and content of document elements. The tactic level comes afterwards, whereby the syntax and words phrasing the content of the document plan are selected. Elements in the database have a generic identifier which can be used to pull out the content.

Departing from the source DTD in Spanish, institutional writers have a document scheme containing either the content of some of the elements or optional elements to choose from, in case there are more than one solution. These elements are the translation units.

Only those document segments that wear a generic identifier and whose contents have been introduced in the translation memory may be automatically translated. If a segment has not been stored in the translation memory, it cannot be translated.

8 Conclusions

This paper has shown how bilingual documentation within specialised domains can be efficiently managed by means of rich mark-up. Complex tags have been introduced in the corpus thereby increasing the value of the annotation scheme. Value added tags have served a wide variety of functions: text segmentation into translation units, bitext alignment, DTD abstraction, translation memory indexation, text retrieval, and DTD-directed document generation.

9 Acknowledgements

This research is being partially supported by the Spanish Research Agency, project ITEM, TIC-96-1243-C03-01.

References

Abaitua, J., Casillas, A. and Martínez, R. (1997). Segmentación de corpus paralelos para memorias de traducción *Procesamiento del Lenguaje Natural* 1997.

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

Ahonen, H. (1995). Automatic Generation of SGML Content Models *Electronic Publishing* 8(2-3), 195-206, 1995

Brown, P., Della Pietra, V., Della Pietra, S., Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19(2):263-301 1993.

Catizone, R., Russell, G., Warwick, S. (1993). Deriving Translation Data from Bilingual Texts. *Proccedings of the First International Lexical Acquisition Workshop,* Detroit, MI, 1993.

Chang, J. S., Chen, M. H. (1997). An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 297-304, 1997.

Collins, B., Cunningham, P., Veale, T. (1996). An Example Based Approach to Machine Translation. *Expanding MT Horizonts: Proceedings of the Second Conference of the Association for Machine Translation in the Americas:AMTA-96,* 125-134, 1996.

Daille, B., Gaussier, E., Lange, J.M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the 15th International Conference on Computational Linguistics,* 515-521, Kyoto, Japan.

Dagan, I., Church, K. (1994). Termigh: Identifying and translating Technical Terminology. *Proceedings Fourth Conference on Applied Natural Language Processing (ANLP-94),* Stuttgart, Germany, 34-40, 1994. Association for Computational Linguistics.

Eijk, P. van der. (1993). 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.

Frantzi, K. T., Ananiadou, S. (1996). Extracting Nested Collocations. *NLP+IA96/TAL+AI96.* Moncton, Canada, 93-98, 1996.

Gale, W., Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts. *Proceedings of the DARPA SNL Workshop,* 1991.

Gale, W., Church, K. W., Yarowsky, D. (1992). Using Bilingual Materials to Develop Word Sense Disambiguation Methods. *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-92), 101-112, Montreal, Canada 1992.

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

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. *Proceedings of the 31st Annual Meeting of the ACL,* Columbus, Ohio, 17-22. Association for Computational Linguistics 1993.

Martínez, R., Casillas, A., Abaitua, J. (1997). 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.

MtSeg: overview. (1997). *Multext - Document MSG 1. MtSeg/Overview*http://www.lpl.univ-aix.fr/projects/multext/MUL7.html.

Shafer, K. (1995). Automatic DTD creation via the GB-Engine and Fred. *http://www.oclc.org/fred/docs/papers/*, 1995.

SIGIR. (1996). *Workshop on Cross-linguistic Multilingual Information Retrieval,* Zurich, 1996.

Smadja, F., McKeown, K., Hatzivassiloglou, V.(1996). Translat-