Encoding a Parallel Corpus for Automatic Terminology Extraction

Johann Gamper
European Academy Bolzano/Bozen
Weggensteinstr. 12/A, 39100 Bolzano/Bozen, Italy
jgamper@eurac.edu

Abstract

We present a status report about an ongoing research project in the field of (semi-)automatic terminology acquisition at the European Academy Bolzano. The main focus will be on encoding a text corpus, which serves as a basis for applying term extraction programs.

1 Introduction

Text corpora are valuable resources in all areas dealing with natural language processing in one form or another. Terminology is one of these fields, where researchers explore domain-specific language material to investigate terminological issues. The manual acquisition of terminological data from text material is a very work-intensive and error-prone task. Recent advances in automatic corpus analysis favored a modern form of terminology acquisition: (1) a corpus is a collection of language material in machine-readable form and (2) computer programs scan the corpus for terminologically relevant information and generate lists of term candidates which have to be post-edited by humans. The following project CATEx adopts this approach.

2 The CATEx Project

Due to the equal status of the Italian and the German language in South Tyrol, legal and administrative documents have to be written in both languages. A prerequisite for high quality translations is a consistent and comprehensive bilingual terminology, which also forms the basis for an independent German legal language which reflects the Italian legislation. The first systematic effort in this direction was initiated a few years ago at the European Academy Bolzano/Bozen with the goal to compile an Italian/German legal and administrative terminology for South Tyrol.

The CATEx (Computer Assisted Terminology Extraction) project emerged from the need to support and improve, both qualitatively and quantitatively, the manual acquisition of terminological data. Thus, the main objective of CATEx is the development of a computational framework for (semi-)automatic terminology acquisition, which consists of four modules: a parallel text corpus, term-extraction programs, a term bank linked to the text corpus, and a user-interface for browsing the corpus and the term bank.

3 Building a Parallel Text Corpus

Building the text corpus comprises the following tasks: corpus design, preprocessing, encoding primary data, and encoding linguistic information.

3.1 Corpus Design and Preprocessing

Corpus design selects a collection of texts which should be included in the corpus. An important criteria is that the texts represent a realistic model of the language to be studied (Bowker, 1996). In its current form, our corpus contains only one sort of texts, namely the bilingual version of Italian laws such as the Civil Code. A particular feature of our corpus, which contains both German and Italian translations, is the structural equivalence of the original text and its translation down to the sentence level, i.e. each sentence in the original text has a corresponding one in the translation. The corpus is one of the largest special language corpora. It contains ca. 5 Mio. words and 35,898 (66,934) different Italian (German) word forms.

In the *preprocessing* phase we correct (mainly OCR) errors in the raw text material and produce a unified electronic version in such a way as to simplify the programs for consequent annotation.

3.2 Encoding Primary Data and Linguistic Annotation

Corpus encoding successively enriches the raw text material with explicitly encoded information. We apply the Corpus Encoding Standard (CES), which is an application of SGML and provides guidelines for encoding corpora that are used in language engineering applications (Ide et al., 1996). CES distinguishes primary data (raw text material) and linguistic annotation (information resulting from linguistic analyses of the raw texts).

Primary data encoding covers the markup of relevant objects in the raw text material. It comprises documentation information (bibliographic information, etc.) and structural information (sections, lists, footnotes, references, etc.). These pieces of information are required to automatically extract the source of terms, e.g. "Codice Civile, art. 12". Structural information helps also to browse the corpus; this is important in our case, since the corpus will be linked to the terminological database.

Encoding linguistic annotation enriches the primary data with information which results from linguistic analyses of these data. We consider the segmentation of texts into sentences and words, the assignment/disambiguation of lemmas and part-of-speech (POS) tags, and word alignment. Due to the structural equivalence of our parallel texts, we can easily build a perfectly sentence-aligned corpus which is useful for word alignment. The above mentioned linguistic information is required for term extraction, which is mainly inspired by the work in (Dagan and Church, 1997). The monolingual recognition of terms is based on POS patterns which characterize valid terms and the recognition of translation equivalents is based on bilingual word alignment. Lemmas abstract from singular/plural variations, which is useful for alignment and term recognition.

For the linguistic annotation we use the MULTEXT tools available from http://www.lpl.univ-aix.fr/projects/multext. We already have extensive experience with the tokenizer MtSeg which distinguishes 11 classes of tokens, such as abbreviations, dates, various punctuations, etc. The customization of MtSeg via language-specific resource files has been done in a bootstrapping process similar to the filter programs. An evaluation of 10% of the Civil Code ($\approx$ 28,000 words) revealed only one type of tokenization error: a full stop that is not part of an abbreviation and is followed by an uppercase letter is recognized as end-of-sentence marker, e.g. in "6. Absatz". This kind of error is unavoidable in German if we refuse to mark such patterns as compounds.

Currently we are preparing the lemmatization and the POS tagging by using MtLex. MtLex is equipped with an Italian and a German lexicon which contain 138,823 and 51,010 different word forms respectively. To include the 15,013 (58,217) new Italian (German) word forms in our corpus the corresponding lexicons have been extended. The creation of the Italian lexicon took 2 MM.

4 Discussion

The general approach we adopted in the preprocessing and primary data encoding phases was to pass the raw texts through a sequence of filters. Each filter adds some small pieces of new information and writes a logfile in case of doubt. The output and the logfile in turn are used to improve the filter programs in order to minimize manual post-editing. This modular bootstrapping approach has advantages over huge parameterizable programs: filters are relatively simple and can be partially reused or easily adapted for texts with different formats; tuning the filters becomes less complex; when recovering from a previous stage the loss of work is minimized. The filters have been implemented in Perl which, due to its pattern matching mechanism via regular expressions, is a very powerful language for such applications.

Future work will include the completion of the linguistic annotation. The MULTEXT tagger Mt-Tag will be used for the disambiguation of POS tags. Word alignment still requires the study of various approaches, e.g. (Dagan et al., 1993; Melamed, 1997). Finally, we are working on a sophisticated interface to navigate through parallel documents to disseminate the text corpus before terminology extraction has been completed.

References

Lynne Bowker. 1996. Towards a corpus-based approach to terminography. *Terminology*, 3(1):27–52.

Ido Dagan and Kenneth W. Church. 1997. *Termight*: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12:89–107.

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8.

Nancy Ide, Greg Priest-Dorman, and Jean Véronis. 1996. Corpus encoding standard. See http://www.cs.vassar.edu/CES/.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of ACL/EACL-97*, pages 302–312.