The English-Romanian parallel corpus NAACL 2003

Table of Contents

1 BASIC INFORMATION

1.1 Corpus composition

The English-Romanian parallel corpus NAACL 2003 supplied in Batch 2 of the MetaNet4U project consists of the News part of the word alignment traning corpus from the HLT-NAACL 2003 workshop "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond" (Mihalcea and Pedersen, 2003). It was collected in 2003, from Romanian newspapers archives with English versions of the published articles. It was automatically sentence aligned and then manual validation was performed on the resulting sentence alignments.

1.2 Representation of the corpus

The corpus is encoded in the XCES XML format (http://www.xces.org/), with one XML file per language.

1.3 Character encoding

The characters are UTF-8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Dan Tufiş
Address: Calea 13 Septembrie, no. 13, 050711
Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
Position: Director
Telephone: +40 21 3188103
Fax: +40 21 3188142
e-mail: tufis@racai.ro

2.2 Delivery medium

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

The corpus consists of two XML files: ‘`NAACL_news-en.xml`’ and ‘`NAACL_news-ro.xml`’.

3.2 Data structure of an entry

The corpus is structured in paragraphs, divided into sentences. Each sentence is segmented into tokens, including punctuation. Each token has a descriptor attribute (‘`msd`’) containing syntactic information about its grammatical category and its morpho-lexical attributes, and a base form attribute (‘`base`’) containing the lemma.

3.3 Corpus size

The corpus contains 39956 sentence pairs. There are 843,832 words (no punctuation) in English and 757,550 words in Romanian. The English file (XML encoded) has 58.4MB and the Romanian file has 54.6MB.

4 CONTENT INFORMATION

4.1 Type of the corpus

Parallel corpus, written, annotated, XCES-encoded.

4.2 The natural language(s) of the corpus

English and Romanian.

4.3 Domain(s)/register(s) of the corpus

The corpus is automatically collected from the web, from Romanian newspapers archives having English translations. The domain of the corpus is “News” with subdomain “Political News”.

4.4 Annotations in the corpus

4.4.1 Types of annotations

The corpus is sentence split, tokenized, POS-tagged (with the MSD tagset in both English and Romanian, see http://nl.ijs.si/ME/V3/msd/msd.pdf) and lemmatized. **The POS tagging for both English and Romanian was checked by hand at the Research Institute for Artificial Intelligence of the Romanian Academy** and thus this is a manually validated POS-tagged corpus. It is suitable for training POS tagging models.

4.4.2 Tags

We used the Multext-East Morpho-Syntactic Descriptors to POS tag this corpus. For further information about MSDs, one can refer to http://nl.ijs.si/ME/V3/msd/msd.pdf.

4.4.3 Alignment information

The corpus contains 39956 sentence pairs. The alignment of the sentences is implicit, i.e. there is no external alignment file.

Every sentence (e.g. ‘`<xces:s id="NAACL_2003_en_1">…</xces:s>`’) in the English file has a unique integer identifier (1) which corresponds to the parallel sentence in the Romanian file (‘`<xces:s id="NAACL_2003_ro_1">…</xces:s>`’).

4.4.4 Attributes and their values

Each token has three attributes: ‘`type`’ which can be ‘word’ or ‘punctuation’, ‘`base`’ which contains the lemma of the token (if the token is of type ‘punctuation’, then the value of the base attribute is the punctuation string) and ‘`msd`’ which contains the correct MSD of that token in context.

4.5 Intended application of the corpus

Being manually validated, the corpus may be used to train POS tagging/lemmatization statistical models. It can also be used to extract English-Romanian translation equivalents and to study word alignment algorithms. It is too small for being useful in deriving n-gram langauge models but parts of it can constitute test/development sets for SMT.

4.6 Reliability of the annotations

The POS annotations are reliable in that the POS tagging has been checked and corrected by hand. Initially, we ran the POS tagging training and testing on this corpus for both English and Romanian doing POS label corrections where necessary, until the tagger's (a variant of the TnT POS tagger (Brants, 2000) called TTL (Tufiș et al., 2008)) accuracy was above 99%. After that, we manually checked every pair of sentences in order to see if the POS tagging is correct.

Lemmatization is also reliable in that every lemma is assigned from a lexicon containing for every word form, its correct lemma and MSD label. The ambiguity cases where for the same MSD label we had different lemmas, were resolved by hand.

5 REFERENCES

Brants, T. (2000). TnT – A Statistical Part-Of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference ANLP-2000*. Seattle, WA, pp 224--231.

Mihalcea, R. and Pedersen, T. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT/NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May 2003.

Tufiș , D., Ion, R., Ceaușu, A., and Ştefănescu, D. RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference* - LREC 2008, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.