YAWA, a Romanian to English Word Aligner

Radu ION

"Mihai Drăgănescu" Research Institute for Artificial Intelligence
Romanian Academy
radu@racai.ro

Table of Contents

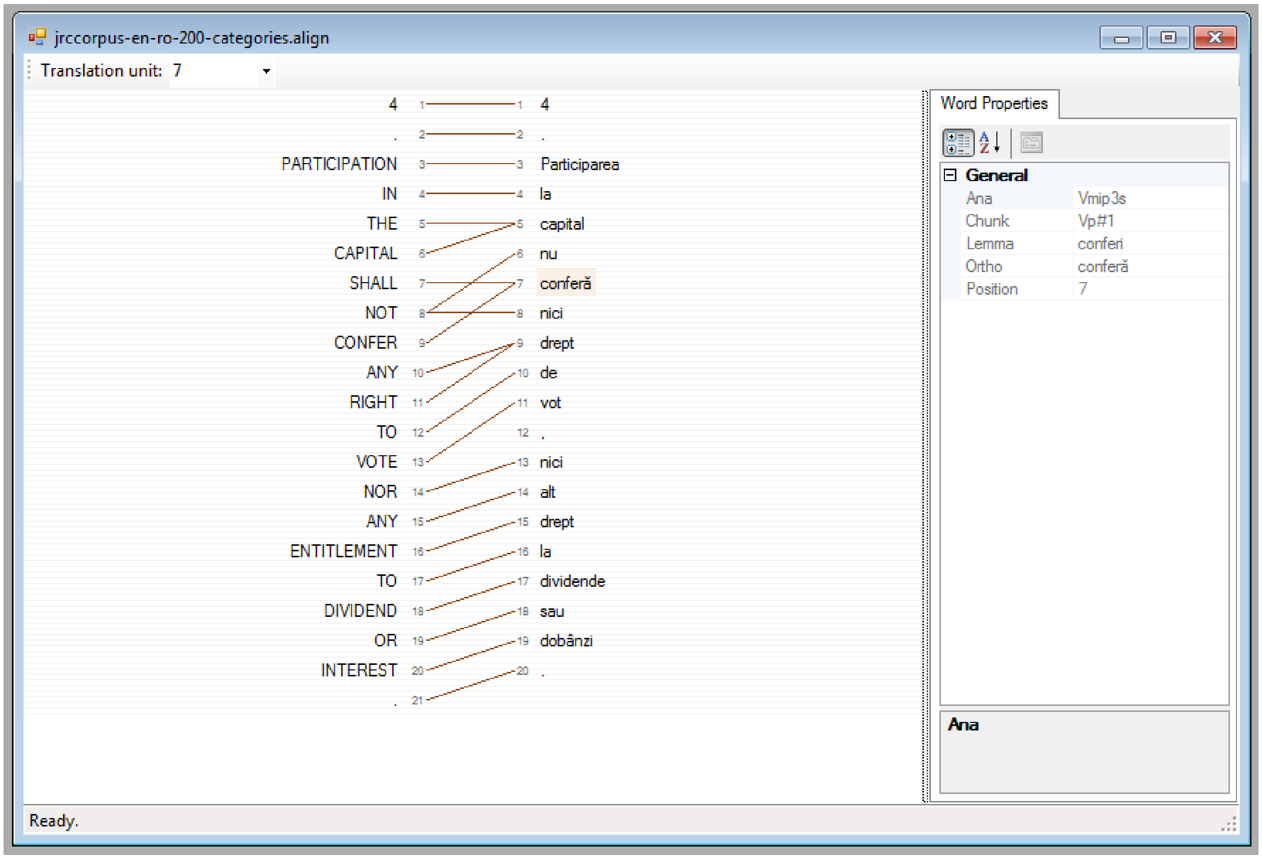
1 BASIC INFORMATION

1.1 Tool name

“Yet Another Word Alignment algorithm” or YAWA for short.

1.2 Overview and purpose of the tool

YAWA is a word alignment tool written in Perl for Romanian-English that generates alignments N:M alignments between words in the source and target sentences. It has the same functionality as the statistical word alignment package GIZA++ (Och and Ney, 2003). The next figure displays a word alignment that YAWA produces.



Translation unit: 7

General

Ana	Vmip3s
Chunk	Vp#1
Lemma	conferi
Ortho	conferă
Position	7

Ana

Ready.

Figure 1: A YAWA word alignment example

1.3 A short description of the algorithm

YAWA is a 4 stage word aligner that complements skeleton alignments produced in a previous stage with additional alignments based on linguistic-driven heuristics. In the last stage, it performs a cleanup of the alignments such that certain restrictions are met (long crossing links are removed, “parallel” links are enforced, etc.) A more detailed description is given in Tufiș et al. (2006).

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

YAWA is written in Perl and it is at version 3.6. The Romanian-English plugin (the configuration file as well as the specific linguistic routines; Perl package) is at version 1.5. Thus, the only requirement is to have a recent version of Perl installed.

2.2 Installation

No installation is required.

2.3 Execution instructions

In order to run YAWA on a specific XML corpus (see the next section for details on the format and annotations) one has to:

- Edit the plugin file ‘`yawaplugin.pm`’ and enter the path to the XML corpus in function ‘`corpusFile()`’ currently at line 66 in this file;
- Run the command: ‘`yawa.pl >alignment.txt`’ from the directory containing the ‘`yawa.pl`’ script. The output of the alignment process is written in the ‘`alignment.txt`’ file.

In order to run, YAWA needs an English-Romanian dictionary (currently located in ‘`dicts/reference.treq`’; see the function ‘`treqDict()`’ from the plugin file). The format of this dictionary is exemplified below:

```
intersect,1+     intersec&tcedil;ie,1+ 1+:63478454      11.7668014876134
…
```

in which, the first four fields (TAB separated) are: source word, comma, the metacategory of this word (a collection of POS classes), TAB, target word, comma, the metacategory of this word, TAB, metacategory, double colon, number of examples, TAB, loglikelihood score (the higher, the more confident we are that the translation pair is correct) and some other fields.

2.4 Input/Output data formats

The input file to YAWA is an XML-encoded corpus that is POS-tagged (with metacategory annotation), lemmatized and chunked in both English and Romanian. The format of the corpus is exemplified by the test file located in the directory ‘`test/jrccorpus-en-ro-200-categories.xml`’. To obtain metacategories over POS tags, a helper script has been provided in the ‘scripts/’ directory (to see what it produces, simply run it on the test file ‘`test/jrccorpus-en-ro-200-no-categories.xml`’).

One can obtain the required annotations by using TTL, another MetaShare4U deliverable which is located at: http://ws.racai.ro:9191/repository/browse/tokenizing-tagging-lemmatizing-and-chunking-free-running-texts/e8f4fe8ed58b11e1a3cb00226410db013c013c214f9a4de0a3bd54d88cef4ca3/. It is the user’s responsibility to convert from TTL’s output to the XML format required by YAWA.

Alternatively, the TTL web service (located at http://ws.racai.ro/ttlws.wsdl) has a function XCES which will perform the conversion on a sentence by sentence basis.

YAWA will output word alignments in the following format (example):

```
1 9 8 S
1 7 7 S
1 5 5 S
1 4 3 S
1 1 1 S
1 8 7 S
1 3 2 S
1 6 6 S
1 0 4 S
1 2 0 S
```

where in the first field (separated by white space) is the translation unit id (of the <tu> element in the XML corpus) followed by the 1-based numbering of word (<w>) and punctuation (<c>) elements in the English (second field) and Romanian (third field) sentences (<s>) in each translation unit. ‘S’ indicates a sure alignment (all alignments are sure). ‘0’ indicates a NULL alignment (or no alignment) for the corresponding index.

2.5 Integration with external tools

Other than running the ‘`scripts/xceswithcats.pl`’ to obtain metacategories over POS tags of the input XML corpus, no other tool is required/called by YAWA.

3 CONTENT INFORMATION

3.1 A test input file

See the file ‘`test/jrccorpus-en-ro-200-categories.xml`’.

3.2 The output file

See the file ‘`test/jrccorpus-en-ro-200-categories.align`’.

3.3 Running time

The test file was processed in approximately 35 seconds on an Intel(R) Core(TM) i7 CPU 980 @ 3.33GHz. The test set contains 200 parallel sentences.

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further details and assistance, please contact the developer: Radu ION, radu@racai.ro.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

Dan Tufiș, Radu Ion, Alexandru Ceaușu, and Dan Ștefănescu. *Improved Lexical Alignment by Combining Multiple Reified Alignments*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pages 153–160, Trento, Italy, April 3–7 2006.