Transferring Coreference Chains through Word Alignment

Oana Postolache*§, Dan Cristea§† and Constantin Orasan‡

*University of Saarland, Saarbrucken, Germany
oana@coli.uni-saarland.de
§ "Alexandru Ioan Cuza" University, Iaşi, Romania
†Institute of Computer Science, Romanian Academy, Iaşi, Romania
dcristea@infoiasi.ro
‡University of Wolverhampton, Wolverhampton, United Kingdom
C.Orasan@wlv.ac.uk

Abstract

This paper investigates the problem of automatically annotating resources with NP coreference information using a parallel corpus, English-Romanian, in order to transfer, through word alignment, coreference chains from the English part to the Romanian part of the corpus. The results show that we can detect Romanian referential expressions and coreference chains with over 80% F-measure, thus using our method as a preprocessing step followed by manual correction as part of an annotation effort for creating a large Romanian corpus with coreference information is worthwhile.

1. Introduction

Annotated corpora are necessary for a wide range of tasks in computational linguistics, but large scale corpora annotated with syntactic, semantic or discourse information are available for only a handful of languages. Recently, many researchers have started to focus on methods for creation of annotated corpora for less wide-spread languages using parallel texts to project annotations across languages. For example, Yarowsky and Ngai (2001) obtain 76% accuracy in projecting POS tags from English to French and Hwa et al. (1994) report 65.7% accuracy in projecting dependency structures from English to Spanish. At the semantic level, Pado and Lapata (2005) propose a framework for projecting semantic roles from English to German for FrameNet annotation and report 65% F-measure for predicate pairs with matching frame assignments.

Coreference resolution is an important subtask which is required in many NLP applications, including information extraction, question answering and automatic summarization (Postolache, 2004; Postolache and Forăscu, 2004). In the field of coreference resolution there is much less work on how parallel corpora can be used and to the best of our knowledge there has been no attempt to project the coreference relations in a different language using parallel corpora as a way of developing resources. Harabagiu and Maiorano (2000) show how the results of a coreference resolver trained on a parallel corpus for English and Romanian outperforms the results of the resolvers trained on each individual corpus. Their explanation for these results is that by using a parallel corpus it is possible to derive more powerful heuristics which are not available in each individual language. In a similar research, Mitkov and Barbu (2000) show that it is possible to improve the results of knowledge-poor anaphora resolution methods by using a parallel corpus. As in the case of (Harabagiu and Maiorano, 2000) the results of the system which uses the parallel corpus are higher than those obtained by monolingual systems. Salmon-Alt and Vieira (2002) analyze whether the rule-based coreference resolver developed for English in (Vieira and Poesio, 2000) can be applied to other languages. To this end, a French-Portuguese parallel corpus is used for evaluation, but no attempt was made to enhance the results using the fact that the corpus is parallel. Instead, the parallel corpus was used to test linguistic hypotheses.

The best coreference resolution systems are based on machine learning approaches, thus, they need a lot of data for training. For languages such as English, annotated corpora exists, however very often this is not the case for other languages.

In this paper we present a method for projecting coreference chains from English to Romanian using a parallel corpus. Our goal is to use this method as a preprocessing step followed by manual correction in an annotation effort for creating a large Romanian corpus. Thus, we aim to have a high precision method so that the annotators don't need to correct much but focus on adding new annotation that the preprocessing phase did not detect.

The rest of the paper is structured as follows. Section 2 describes the parallel corpus we are using, Section 3 presents the experiments and Section 4 shows the results. In Section 5 we do the error analysis and we conclude in Section 6.

2. Parallel coreference corpus

The experimental corpus consists of three parts from the first chapter of the English original and Romanian translation of the novel "1984"[1].

The data is annotated with coreference information. We consider any prototypical noun phrases and noun phrase surrogates that have a referential function in the sense that they are used to either introduce a new entity in the discourse or to refer to an already existing one. We use the term referential[2] expression (RE) for all these noun phrases.

[1] George Orwell, "1984". Secker and Warburg, 1949. Romanian translation by Mihnea Gafiţa: "O mie nouă sute optzeci şi patru". Editura Univers, Bucharest.

[2] We use the term 'referential' to mean that the expression refers to a physical/abstract entity in the discourse, rather than to an expression appearing earlier in the text. In this sense, an indefinite noun-phrase is an RE.

	Text 1		Text 2		Text 3		Total	
	En	Ro	En	Ro	En	Ro	En	Ro
Words	6,826	6,952	3,220	3,220	2,871	2,961	12,917	13,133
Sentences	296	296	178	178	164	164	638	638
REs	1,866	1,726	869	845	875	851	3,610	3,422
DEs	926	950	467	516	425	481	1,818	1,947

Table 1: Statistics of the experimental data.

Coreference annotation involves determining whether or not two REs are used in the text to refer to the same entity. Coreference is therefore an equivalence relation that groups the REs in a text in equivalence classes. We name these equivalence classes discourse entities (DEs).

The English part of the corpus was processed using an FDG-tagger (Järvinen and Tapanainen, 1997) which provided POS information and dependency relations. The noun-phrases and their heads were then automatically identified (i.e., all the structures dominated by a head noun or pronoun were considered NPs). Human annotators then manually eliminated the errors from the automatic detection of the NPs and marked new REs, other than NPs or pronouns. Our referential expressions are generally conformant with the MUC-7 (Hirschman and Chinchor, 1997) and ACE (2003) criteria, although there are some differences.

The types of mentions that we consider as REs are: **noun phrases**: definite (*the shut window-pane*), indefinite (*a bright cold day*) or undetermined (*sole guardian of truth*); **names** (*Winston Smith*, *The Ministry of Love*); **pronouns**: personal, possessive, reflexive and demonstrative; **wh-pronouns**: relative pronouns (*which*, *who*, etc.); **numerals**, when they refer to entities (*the first*).

It is important to note that our REs include only the restrictive relative clauses, each term of an apposition is taken separately ([*Big Brother*], [*the primal traitor*],), conjoined expressions are annotated individually ([*John*] and [*Mary*]), and noun premodifiers are not marked ([*glass doors*]).

The Romanian part of the corpus was manually annotated for referential expressions, their heads, and coreference chains. We have used the same guidelines as for English at which we added rules for some language specific cases. This Romanian annotation was used for evaluation purposes. Table 1 shows statistics about the experimental corpus.

3. Experiments

The work presented in this paper involves three steps:

1. Automatic word alignment using a Romanian-English aligner. We have used the COWAL Romanian-English aligner (Tufiş et al., 2006), that has a performance of 83.30% F-measure. A previous version of the system (Tufiş et al., 2005) participated in the ACL2005 shared task on Word Alignment, and was ranked first out of 37 competing systems.

2. Extraction of Romanian REs corresponding to English REs. For each English RE which spans *n* words e_1, e_2, ..., e_n, we extract the corresponding set of Romanian words, with which the English words are aligned, r_{i_1}, r_{i_2}, ..., r_{i_n}. We order the Romanian words according to the surface order, remove duplicates, and consider the corresponding Romanian RE as the span of words between the first and the last word. We mark as the head of the resulting RE the Romanian word(s) aligned with the head of the English RE. Because the word alignment is $n : m$, where $n, m \geq 0$, we can encounter the following situations:

(a) An English RE has a corresponding Romanian RE with one head.

(b) An English RE has a corresponding Romanian RE with two or more heads (when the English head is aligned with more than one Romanian word).

(c) An English RE has a corresponding Romanian RE with no head (when the English head is not aligned with any Romanian word).

(d) An English RE has no corresponding Romanian RE (when no words of the English RE are aligned with any Romanian word).

Among the four situations above we only consider Romanian REs that have at least one head (situation (a) and (b)). We choose as the head of the Romanian RE the leftmost Romanian word whose part of speech is Noun, Pronoun or Numeral (if none of the Romanian words corresponding to the English head have one of these POS tags, the RE is discarded).

Table 3 shows the distribution of the transferred REs among the four situations mentioned above, the number of transferred REs for which the head had a wrong POS tag and the number of final REs considered.

3. Transfer of English coreference chains to Romanian. As the English REs are clustered in groups referring to the same entity, and as we have the corresponding Romanian REs, we simply 'import' the clustering. As we have seen above, not all the English REs have a corresponding Romanian RE, which triggers different numbers of groups (DEs) in the English part and the Romanian (transferred) part. Table 2 shows the number English DEs and the corresponding transferred Romanian ones.

	Text 1	Text 2	Text 3	Total
En DEs	926	467	425	1,818
Ro DEs	778	401	373	1,552

Table 2: The English discourse entities (DEs) and the corresponding transferred Romanian ones

	Text 1	Text 2	Text 3	Total
English REs	1,866	869	875	3,610
Romanian REs - one head	1,247	592	590	2,429
Romanian REs - more heads	102	49	52	203
Romanian REs - no head	68	33	23	124
No Romanian REs	449	195	210	854
Romanian REs - heads with wrong POS	82	35	24	141
Final Romanian REs	1,267	606	618	2,491

Table 3: Romanian REs resulted through the word alignment transfer

4. Evaluation

The previous three steps lead to automatic detection of REs and coreference chains for the Romanian part of the parallel corpus. We evaluate the transferred REs and the induced coreference chains (we call them system) against the manually annotated Romanian corpus (gold standard). We perform three types of evaluation:

1. Evaluation of the RE heads

We consider only the heads of the system REs and the heads of the gold standard REs and compute Precision, Recall and F-measure. The values are shown in Table 4.

	Text 1	Text 2	Text 3
Precision	95.81	95.37	95.95
Recall	70.33	68.40	69.68
F-measure	81.12	79.66	80.73

Table 4: Evaluation of the RE heads

2. Evaluation of the RE spans

Instead of looking only at heads, we consider the overlap between the system REs and the gold standard REs. When computing Precision and Recall, in the numerator instead of counting how many correct heads the system has found, we sum the overlaps between the system REs and the gold standard REs. The overlap is computed as twice the number of words in the intersection of the two REs devided by the sum of the number of words in the two REs.

We do this type of evaluation in two ways. First, we evaluate the system REs against all reference REs. Table 5 shows the Precision, Recall and F-measure for this evaluation. Then, because the numbers we obtain in this manner also include the penalties for not having certain REs in the system, or having some wrong REs (errors that are shown through the evaluation of RE heads above), we computed a score that shows, only for the correct REs found by the system (that is, with correct heads), what is the accuracy of the span detection. Thus, we evaluated the correct system REs against the corresponding gold standard REs. The values, called Span Overlaps, are shown in the last line of Table 5.

3. Evaluation of coreference chains.

As coreference evaluation is a controversial issue, we perform the evaluation using two metrics: the MUC-score (Vilain et al., 1995) which considers only DEs with more than one RE and the B-cubed score (Bagga and Baldwin, 1998) that also considers DEs with single REs.

As in the case of RE span evaluation, we also evaluate coreference chains in two ways. First, we evaluate all system REs and coreference chains against all gold standard REs and coreference chains. The results are shown in Table 6. Then, in order to eliminate the errors that stem from incorrect detection of REs, and to get an idea of how well the coreference transfer worked by itself, we evaluate the correct system REs and their transferred chains against the corresponding gold standard REs and chains. The results are shown in Table 7.

	Text 1	Text 2	Text 3
Precision	86.58	86.03	87.92
Recall	63.55	61.70	63.85
F-measure	73.30	71.86	73.97
Span Overlap	90.36	90.20	91.63

Table 5: Evaluation of RE spans

		Text 1	Text 2	Text 3
MUC score	P	53.35	52.27	51.35
	R	84.66	83.90	77.55
	F	65.45	64.41	61.78
B-cubed score	P	72.95	76.05	72.24
	R	94.35	95.40	91.81
	F	82.29	84.64	80.86

Table 6: Coreference chains evaluation using all system and gold standard REs

		Text 1	Text 2	Text 3
MUC score	P	88.46	90.52	90.04
	R	89.80	86.86	81.19
	F	89.12	88.65	85.39
B-cubed score	P	92.67	95.33	94.43
	R	93.30	93.95	89.34
	F	92.99	94.63	91.81

Table 7: Coreference chains evaluation using only the correct system REs and coresponding gold standard REs

5. Error Analysis

The first set of errors (that also propagate throughout the whole process) come from the incorrect detection of Romanian REs. The recall is about 70%, while the precision is quite high (above 95%). We analyzed the reasons for the low recall, and found four types of errors (in what follows, the REs in boldface don't have a corresponding RE in English, so they couldn't be detected):

1. Wrong alignment: English heads that are either not aligned with any Romanian word or with wrong ones.
2. English adjectives or adverbs translated in Romanian as PPs which include NPs, e.g., *naturally sanguine face* is translated as *faţă sangvină de la* ***natură*** (face sanguine from **the nature**). Also, there are cases in which English verbs are translated in Romanian as NPs: *To mark the paper was the decisive act* vs. ***Datarea hârtiei*** *era actul decisiv* (**The 'date marking' of the paper** was the decisive act).
3. Additions of the Romanian translator: *The actual writing would be easy* is translated as *Scrisul in sine era* ***o treabă uşoară*** (The writing itself was **an easy job**).
4. English noun premodifiers translated in Romanian as noun postmodifiers or possessives, e.g., *a forced labour camp* is translated as *un lagăr de* ***muncă silnică*** (a camp of **labour forced**) or *the lift-shaft—uşa* ***liftului*** (the door of **the lift**).

The second type of errors occurs in the span overlap. Here, most of the errors are due to incorrect alignment, but some are triggered by the translation, e.g., ***Someone with a comb and a piece of toilet paper*** *was trying to keep tune with the music* translated as ***Cineva*** *se străduia, cu un pieptene şi o bucată de hârtie igienică, să ţină isonul muzicii.* (**Somebody** was trying with a comb and a piece of toilet paper to keep tune with the music.).

Finally, we analyzed the errors in the coreference chains detection. Clearly, when evaluating against the whole set of gold standard REs (Table 6) most of the errors are due to the previous step—REs detection. However, when we evaluate the chains considering only the correct system REs, the results are not 100%, as it would be expected. This is due to the translation choice, e.g., in ***The sky*** *was* ***a harsh blue***, translated as ***Cerul*** *era de* ***un albastru strident*** (**The sky** was as **a harsh blue**), the two REs in boldface are coreferent in English, but they are not in the Romanian version.

6. Conclusions

We presented an automatic method for projecting coreference chains in parallel corpora as a preprocessing step prior to manual correction in an annotation effort aiming at creating large scale corpora with coreference information. To illustrate the methodology, a small English-Romanian parallel corpus where each part contained almost 13,000 words has been annotated for coreference. The English set of texts was used as the source of the coreference chains, whilst the Romanian equivalents constituted the gold standard used in the evaluation. The results of the transfer of the referential expressions from English to Romanian show very high precision (over 95%) but lower recall (around 70%). For the transferred coreference chains on correctly detected referential expressions the B-cubed F-measure was over 90% indicating the appropriateness of our method for this task. Error analysis revealed that the low recall stems from referential expressions that could not be aligned to each other due to errors in the automatic word alignment and language differences introduced in the translation.

In the future, we plan to apply our method to larger corpora in order to investigate it further. Even though there are several English-Romanian parallel corpora, in none of them the English texts are annotated with coreference chains. In light of this, we plan to automatically identify coreferential chains in the English texts using a state-of-the-art coreference system, and then transfer them to Romanian. Even though this will probably lead to lower results than when manually annotated corpora are employed, we believe that this approach will still enable us to bootstrap the annotation of coreferential chains in the Romanian section of the parallel corpus. In order to find out the drop in the performance of such an approach, we plan to apply it to the test corpus used in this paper.

Acknowledgements. We thank Dan Tufis and his team for making available to us the Romanian-English aligner.

7. References

ACE. 2003. Entity Detection and Tracking - Phase 1. www.itl.nist.gov/iaui/894.01/tests/ace/phase1/index.htm.

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Conference Workshop at LREC'98*, pages 563–566.

S. M. Harabagiu and S. J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the ANLP-NAACL2000*, Seattle, WA.

L. Hirschman and N. Chinchor. 1997. MUC-7 Coreference Task Definition, version 3.0. In *Proceedings of MUC-7*. See also: http://www.muc.saic.com.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 1994. Bootstrapping parsers via syntactic projection across. *Computational Linguistics*, 20:535–561.

T. Järvinen and P. Tapanainen. 1997. A Dependecy Parser for English. Technical report, University of Helsinky.

R. Mitkov and C. Barbu. 2000. Mutual enhancement of performance: Bilingual pronoun resolution for English and French. In *Proceedings of DAARC2000*.

S. Pado and M. Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of HLT/EMNLP2005*, pages 859–866, Vancouver, Canada.

O. Postolache and C. Forăscu. 2004. A coreference model on excerpt from a novel. In *Proceeding of The European Summer School in Logic Language and Information - ESSLLI'2004*, pages 202–213, Nancy, France.

O. Postolache. 2004. RARE - Robust Anaphora Resolution Engine. Master's thesis, University of Iaşi.

D. Tufiş, R. Ion, A. Ceauşu, and D. Stefănescu. 2005. Combined aligners. In *ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan.

D. Tufiş, R. Ion, A. Ceauşu, and D. Stefănescu. 2006. Improved lexical alignment by combining multiple reified alignments. In *Proceedings of the EACL2006*.

R. Vieira and M. Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539 – 593.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirshman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the MUC6*, pages 45–52.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the NAACL2001*.