Translation- and projection-based unsupervised coreference resolution for Polish*

Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences

Abstract. Creating a coreference resolution tool for a new language is a challenging task due to substantial effort required by development of associated linguistic data, regardless of rule-based or statistical nature of the approach. In this paper, we test the translation- and projection-based method for an inflectional language, evaluate the result on a corpus of general coreference and compare the results with state-of-the-art solutions of this type for other languages.

1 Introduction

A widely known problem of coreference resolution — the process of "determining which NPs in a text or dialogue refer to the same real-world entity" [1], crucial for higher-level NLP applications such as text summarisation, text categorisation and textual entailment — has so far been tackled from many perspectives. However, there still exist languages which do not have state-of-the-art solutions available, which is most likely caused by the substantial effort required by development of language resources and tools, some of them knowledge-intensive, either leading to development of language-specific rules or preparation of training data for statistical approaches.

One of the solutions to this problem is following the translation-projection path, i.e., (1) translating the text (in the *source* language) to be coreferentially annotated into the *target* language, for which coreference resolution tools are available, (2) running the target language coreference resolver, (3) transferring the produced annotations (mentions — discourse world entities and clusters — sets of mentions referring to the same entity) from the target to the source language. Such a solution has so far been proposed e.g. by Rahman and Ng [2] and evaluated for Spanish and Italian with projection from English (see Section 2). Although the source and target languages in this setting come from two different language families, they differ markedly from inflectional languages such as Polish, which makes the approach interesting to test with different language pairs.

* The work reported here was carried out within the *Computer-based methods for coreference resolution in Polish texts (CORE)* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40) and *University Research Program for Google Translate*.

For Polish, there currently exist two resolvers of general coreference, a rule-based [3] and a statistical one [4], yet they were evaluated with a dataset of limited size — unavailable at the time of their preparation. Presently, a new corpus is being built to improve development and evaluation of coreference resolution tools — a Polish Coreference Corpus [5], parts of which have been used to evaluate our experimental results.

2 Related Work

Rahman and Ng's paper refers to many previous projection attempts in NLP tasks, mostly in the context of projecting annotations from a resource-rich to a resource-scarce language, starting from parallel corpus-based solutions to newer, machine translation-based ones. In the context of coreference resolution, two Romanian-English works are mentioned: [6] and [7] and a Portuguese-English one [8], all involving projection of hand-annotated data. Unlike others, Rahman and Ng's approach concentrated on "a technology for projecting annotations that can potentially be deployed across a large number of languages without coreference-annotated data"[1].

The article presents three settings differing in terms of application of linguistic tools, potentially caused by their (un)availability for the source language. Setting 1 assumes no linguistic tools available, which results in projecting not only coreference clusters, but also complete mentions. Setting 2 employs existing mention extractors (as in our case), while setting 3 makes use of all available linguistic processing tools used to generate features and train coreference resolvers on the projected coreference annotation.

As expected, the results of Setting 1 are highly unsatisfactory, with CONLL[2] F1 = 37.6% for Spanish and 21.4% for Italian. Results of setting 2 and 3 show considerable improvement, amounting to 50-60% F-measure.

3 The Experiment

Our experiment concentrated on a configuration combining Rahman and Ng's settings 1 and 2. A Polish text has been translated into English and mentions have been identified in the Polish part (as with setting 2), but an English coreference resolver was running on plain English text — and not on pre-identified Polish mentions transferred to English (as with setting 1). Only then English coreference clusters were used to form Polish clusters using original Polish mentions aligned with English mentions. We believe that this configuration can generally improve translation-based coreference resolution since predetermining mentions might propagate errors resulting e.g. from incorrect classification of nominal constituents of idiomatic expressions as referential. With no mentions predefined,

[1] See [2], bottom of p. 721.
[2] Calculated as (MUC + B^3+ CEAFE) / 3.

the resolver can exclude non-referential expressions in the very first step of the process.

Google Translate service has been used for producing translations, end-to-end coreference baseline system presented in [3] was used for Polish mention detection (see Table 3 for results of mention detection) and Stanford CoreNLP [9], one of the best coreference resolution systems up to date, has been used for English mention detection and coreference resolution. Instead of using external aligners such as GIZA++ [10] employed by Rahman and Ng, we decided to make use of the internal alignment algorithm of Google[3], concentrating the two steps of the process into one, potentially offering better coherence of the result due to internal dependence of both steps — translation and alignment.

Mention statistics	
Gold mentions	23069
Sys mentions	21861
Common mentions	15060

Mention detection results	
Precision	68.89%
Recall	65.28%
F1	67.04%

Table 1. Polish mention detection

Texts for the experiment were acquired from the Polish Coreference Corpus to facilitate evaluation. They constituted 260 gold samples (all currently available), each between 250 and 350 segments, manually annotated with information on mentions and coreference clusters[4].

The following algorithm was used:

Algorithm 1 Translation and projection-based coreference resolution

```
annotate pl-text to detect pl-mentions
translate pl-text into en-text with word-to-word alignment
run en-coreference resolution tool on en-text to detect en-mentions and en-clusters
for all en-clusters (including singletons) do
  for all en-mentions in en-cluster do
    if exists alignment between en-mention head with any pl-mention head then
      put pl-mention in pl-cluster corresponding to en-chain
    end if
  end for
end for
for all pl-mentions not in any pl-cluster do
  create singleton pl-clusters
end for
```

[3] Made available by the University Research Program for Google Translate, see `http://research.google.com/university/translate/`.

[4] See [5], Section 5, for detailed information on organization of the annotation procedure.

4 Evaluation

All usual evaluation metrics have been calculated by comparing projection results with the golden data:

Evaluation metrics	P	R	F
MUC	50.30%	29.62%	37.28%
B^3	93.34%	84.20%	88.53%
CEAFM	81.51%	81.51%	81.51%
CEAFE	81.06%	89.62%	85.12%
BLANC	71.43%	60.51%	64.01%
CONLL	74.90%	67.81%	70.31%

Table 2. Experimental results

The final results show a promising direction and surpass figures given by Rahman and Ng for Spanish and Italian (as compared with best results — except for MUC — in all settings). They even withstand comparison with the official scores of CoNLL-2011 for the top ranked system[5] (below 60% average F1).

The figures could be further improved by investigating how target language-specific properties are being used by the translation-projection process, since inability to fully capture such features is usually considered to be the major weakness of projection-based approaches. However, the commonly cited problematic example of zero pronouns does not hold in the case of languages such as Polish, since their features can easily be propagated onto verbs based on inflectional endings, as in:

(1) *Maria od zawsze kochała Jana. Gdy Øpoprosił ją o rękę, Øbyła szczęśliwa.*[6]
'Maria has always loved John. When he asked her to marry him, she was happy.'

This fact, along with the integration of alignment into translation, might explain the better results for Polish than for Italian or Spanish.

It has also been noticed that the translation-based approach benefited from pragmatic information integrated in the source coreference resolver and propagated without integrating any similar resources into the resolution process for the target language. For example, *atrakcja* 'attraction' mention has been correctly linked by the process with *parada* 'parade' and *miano* 'appellation' with *tytuł* 'title'. This seems to be a very interesting feature since it introduces the idea of exploiting the knowledge used by various coreference resolution tools, also from

[5] See e.g. `http://nlp.stanford.edu/software/dcoref.shtml`.

[6] Actual translation from Polish to English produced by Google Translate, as of March 2013.

different languages. Similarly, a voting mechanism between several target language coreference resolvers could be used in the process to improve the final result.

5 Conclusions and Further Work

We believe that the presented approach can facilitate construction of computational coreference resolvers in two respects: firstly, by creating a useful baseline for languages still lacking coreference resolution tools, and secondly, by applying external knowledge resources to current systems.

A new branch of research could concentrate on the application of different algorithms of alignment of coreference clusters; for languages which have coreference resolvers available, their efficiency could be improved e.g. by testing how corresponding clusters align in the source vs. target language. This could attach singleton mentions in the source language to existing clusters, pointed out by a respective cluster in the target language (i.e. containing a "target" mention aligned with the singleton "source" mention). Investigating how translation quality influences projection results seems another interesting issue.

Last, but not least, the combined translation-alignment procedure could be applied to the data sets used by Rahman and Ng to further improve their results.

References

1. Ng, V.: Supervised Noun Phrase Coreference Research: The First Fifteen Years. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden (2010) 1396–1411
2. Rahman, A., Ng, V.: Translation-Based Projection for Multilingual Coreference Resolution. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2012), Montréal, Canada, Association for Computational Linguistics (2012) 720–730
3. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In Vetulani, Z., ed.: Proceedings of the Fifth Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, Wydawnictwo Poznańskie (2011) 167–171
4. Kopeć, M., Ogrodniczuk, M.: Creating a Coreference Resolution System for Polish. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, ELRA, European Language Resources Association (2012) 192–195
5. Ogrodniczuk, M., Zawisławska, M., Głowińska, K., Savary, A.: Coreference annotation schema for an inflectional language. In Gelbukh, A.F., ed.: Proceedings of the 14th International Conference Computational Linguistics and Intelligent Text Processing (CICLing 2013), Part I. Lecture Notes in Computer Science 7816, Springer (2013) 394–407
6. Harabagiu, S.M., Maiorano, S.J.: Multilingual coreference resolution. In: Proceedings of Sixth Applied Natural Language Processing Conference, North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), Seattle, Washington, USA (2000) 142–149

7. Postolache, O., Cristea, D., Orasan, C.: Transferring Coreference Chains through Word Alignment. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, ELRA, European Language Resources Association (2006) 889–892
8. de Souza, J.G.C., Orasan, C.: Can projected chains in parallel corpora help coreference resolution? In Hendrickx, I., Devi, S.L., Branco, A.H., Mitkov, R., eds.: Anaphora Processing and Applications: Proceedings of the Eighth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Revised Selected Papers. Lecture Notes in Computer Science 7099, Springer (2011) 59–69
9. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics **39**(4) (2013) (forth.)
10. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. ACL 2000, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 440–447