Translation- and Projection-Based Coreference Resolution for Polish

Maciej Ogrodniczuk | Institute of Computer Science
Polish Academy of Sciences

CORE

Language Processing and Intelligent Information Systems
Warsaw, June 18, 2013

Noun phrase coreference resolution

CORE

Two-step process:

1. Identify mentions
2. Build coreference chains with mentions having identical referent

What it really means (here):

1. Mention = NP = a group of adjacent words having nominal head, e.g. pronouns, proper nouns, nominal groups etc.
2. Nesting allowed: *dyrektor departamentu* (EN: *director of the department*)
3. *Identity of reference*

Noun phrase coreference resolution

CORE

Two-step process:

1. Identify mentions
2. Build coreference chains with mentions having identical referent

What it really means (here):

1. Mention = NP = a group of adjacent words having nominal head, e.g. pronouns, proper nouns, nominal groups etc.
2. Nesting allowed: *dyrektor departamentu* (EN: *director of the department*)
3. *Identity of reference*

Why is CR difficult?

CORE

Because it's complex:

Development of associated linguistic data requires substantial effort:

- language-specific rules
- training data for statistical approaches
- knowledge-intensive resources.

But:

While there are no efficient coreference resolution tools for language A ("resource-scarce"), there can be such tools for language B ("resource-rich"), so why not use translation and projection?

Why is CR difficult?

CORE

Because it's complex:

Development of associated linguistic data requires substantial effort:

- language-specific rules
- training data for statistical approaches
- knowledge-intensive resources.

But:

While there are no efficient coreference resolution tools for language A ("resource-scarce"), there can be such tools for language B ("resource-rich"), so why not use translation and projection?

How?

The simpler plan:

A translation/projection-based approach:

- translate the text in A to B,
- resolve coreference in B text using state-of-the art tools,
- transfer the produced annotations from B to A:
  - mentions — discourse world entities
  - clusters — sets of mentions referring to the same entity.

Why not try it for Polish?

CORE

What would we need to do?

- prepare the X-Polish translate-resolve-project tool
- evaluate the result (on a corpus of Polish general coreference)
- compare the results with other solutions of this type for Polish and other languages.

Previous CR projection attempts

CORE

English-Romanian:

- Harabagiu and Maiorano (2000): manual translation of the MUC-6 corpus into Romanian and manual projection of the English annotations to Romanian
- Postolache et al. (2006): automatic word alignment, projection of manual annotations and manual error-fixing.

Different approaches, different goals:

- deep language-related knowledge involved vs. knowledge-lean
- manually annotated data-based vs. fully automatic
- restricted to the given language pair vs. technology applicable to a larger number of languages.

Previous CR projection attempts

English-Romanian:

- Harabagiu and Maiorano (2000): manual translation of the MUC-6 corpus into Romanian and manual projection of the English annotations to Romanian
- Postolache et al. (2006): automatic word alignment, projection of manual annotations and manual error-fixing.

Different approaches, different goals:

- deep language-related knowledge involved vs. knowledge-lean
- manually annotated data-based vs. fully automatic
- restricted to the given language pair vs. technology applicable to a larger number of languages.

Rahman and Ng's solution

Basic assumptions:

- translation with Moses
- alignment with GIZA++
- coreference resolution with Reconcile
- evaluated for Spanish and Italian with projection from English.

Rahman and Ng's solution results

F1 for 3 settings:

1. no linguistic tools available; not only coreference clusters, but also complete mentions are projected: **ES: 37.6%**, **IT: 21.4%**
2. existing mention extractors are employed: **ES: 54.9%**, **IT: 46.8%**
3. all available linguistic processing tools are used to generate features and train coreference resolvers on the projected coreference annotation: **ES: 57.7%**, **IT: 51.7%**.

Non-projection-based state-of the art:

Coreference Resolution in Multiple Languages CoNLL shared task results, 2010: **ES: 60.0%**, **IT: 49.6%**.

Rahman and Ng's solution results

F1 for 3 settings:

1. no linguistic tools available; not only coreference clusters, but also complete mentions are projected: **ES: 37.6%**, **IT: 21.4%**
2. existing mention extractors are employed: **ES: 54.9%**, **IT: 46.8%**
3. all available linguistic processing tools are used to generate features and train coreference resolvers on the projected coreference annotation: **ES: 57.7%**, **IT: 51.7%**.

Non-projection-based state-of the art:

Coreference Resolution in Multiple Languages CoNLL shared task results, 2010: **ES: 60.0%**, **IT: 49.6%**.

The Experiment

Combination of Rahman and Ng's settings 1 and 2 for Polish:

1. Polish text translated into English and mentions identified (as with setting 2)
2. English coreference resolver running on plain English text (not on pre-identified Polish mentions transferred to English as with setting 1)
3. English coreference clusters used to form Polish clusters using original Polish mentions aligned with English mentions.

The Experiment

Reasons for the experiment:

- To test whether it lets avoid errors resulting e.g. from incorrect classification of nominal constituents of idiomatic expressions as referential.
- With no mentions predefined, the resolver can exclude non-referential expressions in the very first step of the process.

System components

Major modules:

1. Google Translate (University Research Program variant):
   - translation
   - word-to-word alignment
2. Polish mention detectors from CORE project:
   - PoliMorf morphological analyser and Pantera tagger for single-word nominal constructs
   - Spejd shallow parser and Spejd grammar of Polish for noun phrases (with nesting and mention boundaries)
   - Nerf for NE recognition
3. Stanford CoreNLP used for English mention detection and coreference resolution.

Why Google Translate?

Two reasons:

1. concentrating the two steps of the process into one
2. offering better coherence of the result due to internal dependence of both steps — translation and alignment.

Resolution algorithm

Translation and projection-based coreference resolution:

```
detect pl-mentions in pl-text
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-cluster
    end if
  end for
end for
for all pl-mentions not in any pl-cluster do
  create singleton pl-clusters
end for
```

Evaluation data

Mentions:

Texts from the Polish Coreference Corpus:

- 260 gold samples (all available at that time)
- each sample between 250 and 350 segments
- manually annotated with information on mentions and coreference clusters.

Mention statistics	
Gold mentions	23069
Sys mentions	21861
Common mentions	15060

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

Experimental results

CORE

Translation- and projection-based approach:

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

Evaluation metrics	P	R	F
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%

Discussion of results

CORE

Two general findings:

- first of all: a useful baseline for languages still lacking coreference resolution tools
- for Polish: the experiment was interesting, but we have better systems now

Further work:

- using the translation-projection method to build coreference resolvers for new languages
- coreference resolution by voting
- testing the approach on Rahman-Ng data set.

Discussion of results

Two general findings:

- first of all: a useful baseline for languages still lacking coreference resolution tools
- for Polish: the experiment was interesting, but we have better systems now

Further work:

- using the translation-projection method to build coreference resolvers for new languages
- coreference resolution by voting
- testing the approach on Rahman-Ng data set.

The CORE project

CORE

Project factsheet:

- Computer-based methods for coreference resolution in Polish texts
- A National Science Centre grant 6505/B/T02/2011/40
- Duration: 2011-2014
- Principal investigator: Maciej Ogrodniczuk

Project summary:

1. Create innovative methods and tools for automated anaphora and coreference resolution in Polish texts
2. Create a corpus of Polish annotated with coreferential chains
3. Test various coreference resolution approaches on the annotated data (rule-based, statistical, hybrid etc.)

Thank you!

It's question time!