Discovery of Common Nominal Facts for Coreference Resolution in Polish

Maciej Ogrodniczuk | Institute of Computer Science
Polish Academy of Sciences

CORE

The 1st International Conference on Mining Intelligence and Knowledge Exploration (MIKE 2013)

Virudhunagar, Tamil Nadu, India, Dec 18, 2013

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.)

Noun phrase coreference resolution

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: $\underline{\textit{director of the }\underline{\textit{department}}}$
3. Identity of reference

Why is CR difficult?

Because it's complex:

Its development requires substantial effort:

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

Why is CR difficult?

CORE

Because it's complex:

Its development requires substantial effort:

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

In either case we make use of:

1. surface features (e.g. linking orthographic entity name with its abbreviation)
2. syntactic features (e.g. traditional gender/number agreement)
3. semantic features (e.g. agreement between semantic classes of mention heads)
4. discourse features (e.g. salience of topics).

Why is it not enough?

Sometimes we have:

- Aldrin and Armstrong stayed friends even though the whole attention of media now focused on the first man on the Moon.
- Mick Jagger appeared on the record of his brother, Chris Jagger. ‘Concertina Jack’ is the 9th album in the younger brother’s discography. The Rolling Stones frontman appeared in ‘Concertina Jack’ and ‘Diamonds are Pearls’ recordings.

Why is it not enough?

Sometimes we have:

- Aldrin and Armstrong stayed friends even though the whole attention of media now focused on the first man on the Moon.
- Mick Jagger appeared on the record of his brother, Chris Jagger. ‘Concertina Jack’ is the 9th album in the younger brother’s discography. The Rolling Stones frontman appeared in ‘Concertina Jack’ and ‘Diamonds are Pearls’ recordings.

Solution:

Using pragmatic features (‘world knowledge’, ‘language use’).

The Nominal Knowledge Base concept

Aren't there existing resources available?

- WordNet? **no**, it does not maintain definitions, cf. *pediatrics* = *branch of medicine that deals with child's diseases*
- isn't just investigating semantic heads of a phrase enough? **no**, there is much difference between *the man* and *the first man on the Moon*

The Nominal Knowledge Base concept

Aren't there existing resources available?

- WordNet? **no**, it does not maintain definitions, cf. *pediatrics* = *branch of medicine that deals with child's diseases*
- isn't just investigating semantic heads of a phrase enough? **no**, there is much difference between *the man* and *the first man on the Moon*

How about:

- extracting definitions from traditional dictionaries (The Dictionary of Periphrastic Constructions, The Great Dictionary of Polish)
- extracting definitions from crowd-sourced dictionaries and bases (`http://sjp.pl`, `http://krzyzowki.info`).

Extraction of collocations

An example:

Krak's (fortified) town is frequently used in media texts about Cracow (for cohesion) — but it will never appear in any ontology.

We will extract such collocations from unstructured sources:

- balanced corpora: the National Corpus of Polish, providing standard representation of a language
- available content sources and electronic media archives: Gutenberg project, Rzeczpospolita Corpus, Polish parliamentary transcripts
- sources of dynamic language: daily electronic media.

and cross-check their occurrence with existing definition bases.

Verification process

Hypothesis:

Pragmatic data available in online data sources could improve coreference resolution in Polish by providing associations unavailable to obtain with currently used methods (surface, syntactic, semantic or discourse-based).

Verification process

Hypothesis:

Pragmatic data available in online data sources could improve coreference resolution in Polish by providing associations unavailable to obtain with currently used methods (surface, syntactic, semantic or discourse-based).

Test setting:

1. extract all coreferential links in the manually annotated set which were missed by the automatic coreference resolver
2. check whether such links could be created if the resolver had access to formalised pragmatic data.

The data set

CORE

Polish Coreference Corpus (short texts):

- 1773 plain texts (31,136 sentences, 503,981 segments)
- 250-350 segments each (284 segments and 18 sentences/text)
- randomly selected from the National Corpus of Polish
- fragments of longer documents
  (but always full consecutive paragraphs)
- manually annotated with mentions and coreference clusters.

The experiment

Data classification:

- 1220 nominal clusters found with differences between manual/automatic annotation
- 73 mention pairs found with data for which coreference resolution was unfeasible with traditional means:
    - 29 personal names linked with person role, function, occupation etc. (e.g. *John Paul II — the Polish pope*, *Rafal Blechacz — a pianist*
    - 18 names of organisations — companies, sports clubs, political parties, music bands etc. (e.g. *Ich Troje — Michal Wisniewski's band*, *Wizzair — low-cost airline*)

The experiment

CORE

Data classification:

- 73 mention pairs found with data for which coreference resolution was unfeasible with traditional means:
  - 14 geographical/geo-political names — here: only names of countries and cities (e.g. *Iraq — the country*, *Aleksandrow Lodzki — the city*)
  - 6 ‘human creation’ names — movie, book and newspaper titles (e.g. *Star Trek — a cinematographic work*, *Foucault's Pendulum — a book*)
  - 6 descriptive definitions, e.g. *cat — domesticated mammal*), *doctors and nurses — hospital staff*).

Knowledge extraction attempt

Sources of pragmatic information:

	Wiki	crossword	both	other	none
personal names	14		14	1	
organisations	9		8		1
geo names		1	13		
creation names	1		5		
definitions	4		1	1	

Data abstraction concept

Expression variation

After extracting the knowledge pieces, different syntax models of a phrase could be built to match the content while maintaining the meaning, e.g.:

- the seed concept: *a prompter*
- variation of participial phrases: *a person who feeds lines to actors — a person feeding lines to actors*
- using wordnet relations to neutralise lexical meaning of phrase components: *a person who feeds lines to performers*.

Summary

CORE

Conclusions:

- currently available data sources can provide pragmatic knowledge
- it can be used to improve coreference resolution in Polish
- the completed version of the database will also find its other linguistic applications:
    - pragmatic analysis of text for smoothing the result of automatic text summarization
    - machine translation
    - readability improvements
- the knowledge base will enrich capabilities of independent IT systems performing text analysis.