Heuristic Detection of Zero Subjects for Polish

Adam Kaczmarek and Michał Marcińczuk

Institute of Informatics
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27,
Wrocław, Poland

November 19, 2014

Politechnika
Wrocławska

Problem definition

Zero subjects

Occurs when a independent clause lacks explicit subject. Omitted subject agreement is reflected by verb morphology.
Toronto Dominion Centre - kompleks handlowo-kulturalny (...).
Ø-Składa się z (...)
Toronto Dominion Centre - trading-cultural complex (...).
Ø-Consists [sg:m:ter] (...)

Zero subject detection

We define the problem of zero subject detection as binary classification of verbs, indicating whether a verb has omitted subject or not.

Introduction

Mention detection

- Detection of zero subjects is considered as part of mention detection
- Mention detection is very important preprocessing step for coreference resolution

Coreference resolution (briefly)

Finding relations between expressions refeering to the same object in real world.

Types of mentions

Mention types in KPWr

In our corpus (KPWr) we distinguish four types of mentions:

- Named entities (United Nations, Google, University of Łódź)
- Agreed phrases (organization, quick brown fox)
- Pronouns (he, she)
- Zero subjects (annotated on verbs)

Motivation

- We tested coreference resolution algorithms on corpus with annotated all verbs and corpus annotated with zero subject detection algorithm. Results indicate that zero subject detection improves significantly results of coreference resolution.

Algorithm name	Annotation type	Precision	Recall	F1
Bartek	all verbs	7.88%	52.21%	13.69%
Bartek	MentionDetector	64.78%	18.59%	**28.89%**
Ikar	all verbs	11.93%	43.52%	18.72%
Ikar	Minos	61.37%	50.16%	**55.20%**
Ruler	all verbs	26.91%	24.54%	25.67%
Ruler	MentionDetector	65.26%	18.21%	**28.48%**

Existing solutions

The problem of zero subject detection and coreference is not very widely studied for Polish.

Mention Detector

- The only existing solution is Mention Detector. Its module for zero subject detection was described and published in spring this year.
- This module implements rule-based zero subject detection based on learning algorithm RIPPER.
- The concepts of **verb** and **noun** were redefined for purposes of this solution

Functional verb classification

Non subject-blocking verbs

These verbs can have zero subject.

Subject-blocking verbs

- lexically:
  - predicatives (pora, sposób)
  - improper verbs (świta, mdli)
- inflexionally:
  - impersonals
  - infinitives
- due to non-prototypic use of person category
  - One should (...) (należałoby, trzeba)
  - Pleonastic "it" (Wczoraj było zimno)
  - One goes/talks/etc. (Idzie się)

Verb and Noun redefined

Analogically to Mention Detector we introduced custom concepts of **verb** and **noun** to better suit problem of zero subject detection.

Verb

Verbs are all words having following PoS': *fin*, *praet*, *bedzie*, *winien*
We excluded some words traditionally considered as verbs *impt*, *imps*, *inf*, *pcon*, *pant*, *pact*, *ppas*, *pred*, *aglt*

Noun

The definition of noun was extended to *gerunds*, *numerals and pronouns*.

Algorithm Overview

Our algorithm combines several methods for classification of **verbs**:

Verb verification

- Rules for discarding verbs
- List of discarded verbs

Subject search

- Dependency parser
- ChunkRel
- Contextual subject search

Subject verification

- Verb-Noun agreement check

Dependency parser and ChunkRel

Dependency parser

- Determines relations between word and its dependents.
- Well suited for free word order languages like Polish.
- MaltParser with model trained on Polish Dependency Bank

ChunkRel

- Determines relations between verbs and its arguments eg. subject and object relations.
- Trained on KPWr

Contextual subject search

To complement relational subject search we performed also contextual subject search limited to window [9, 15] tokens around classified verb.

Allowed cases

nom + **acc** (only for *subst*)

Punctuation and other verbs

Subjects separated by other verbs or by punctuation indicating border of simple sentence within complex sentence are discarded.

Verb-Noun Agreement

Basic agreement

- proper PoS
- proper noun case
- number agreement
- gender agreement
- person agreement

Exceptions

There are some cases, in which we do not want to perform agreement check. One example could be relations from ChunkRel for which experiments shown that gives better results without checking agreement between verb and possible subject.

Verb exceptions

First/Second person

- first/second person explicit pronoun (ja/ty/my/wy)
- please + inf/mr/mrs (Proszę usiąść, Proszę pana)

Third person

- reflexive verbs ending with "ło" (zachmurzyło się)
- auxiliary "it" (zrobiło to ogromne wrażenie)
- preceeding predicative (można było)
- to be + better/worse (jest lepiej niż wczoraj)
- Tak się (tak się składa, że)
- It may be (może być)
- neutral + adv,"już"/num/ (jest już późno, 60 było białych)

Noun exceptions

Noun exceptions

- Numeral with neuter verb - constructions where verb with neutral gender occurs with numeral followed by dative or genitive noun. Example.: 106 posłów głosowało (106 deputies[*gen*] voted[*neutral*])
- Constructions: neg + to be / to have (sic!) + noun[*gen*] Because these verbs are assumed as non-having subject at all
- Adjectives bound as subjects by MaltParser or ChunkRel Sometimes they are part of phrase being subject.

Test data

Our algorithm was developed and tested on Polish Coreference Corpus subset, the same that was used in publication describing MentionDetector module for zero subject detection, to allow direct comparison. (779 documents, 13k sentences, 22k verbs incl. 6k verbs with zero subjects).

Development

Corpus split

Corpus was initially split into two parts:

- development
- test

Development process

The *development* part was used to make observations about verb properties and to verify partial results during development process. After each change error analysis was performed on this part of corpus. The *test* part was used only to calculate the final score.

Results

Algorithm	Precision	Recall	F1
MentionDetector (published)	71.97%	67.39%	69.60%
MentionDetector (unpublished)	**76.45%**	65.14%	70.34%
MINOS	72.43%	**84.72%**	**78.09%**

Conclusions

- Zero subject detection improves significantly coreference resolution
- Heuristic approach performs better
- Lower precision and highest recall are better because of possibility of rejecting verb at level of coreference resolution
- Statistical methods can be better explored and can use richer set of features
- Our algorithm is implemented as part of Liner2 toolkit named MINOS (Mention IdentificatioN for Omitted Subjects)

Future work

- Convert current system into features for statistical learner
- Try deep learning approach

Thank you

Thank you