Semantic Case Role Detection for Information Extraction

Rik De Busser, Roxana Angheluta and Marie-Francine Moens

Introduction

Very roughly, information extraction could be defined as a discipline in which it is attempted to extract semantically relevant elements from a text, using only shallow analysis. This has very often much to do with identifying semantic case roles, i.e. with the detection of the events, states, and their participants as they are mentioned in a text.

Unfortunately, case role detection as a goal in itself has very often been treated in a rather trivial way. Using notions from systemic-functional linguistics, we will try to build a model for extracting generic semantic case roles, i.e. case roles that are not specialized to any particular domain. These roles are learned from a tagged and hand-annotated corpus and can be reassigned to previously unseen text.

Our research will focus on two major questions:

- How shallow can you be? To what extent is there a relationship between the generic semantic plane and morphosyntactic and lexical surface phenomena? What is the minimal amount of information you need?
- Will a shallow approach be useful for information extraction? Can the mapping between morphosyntactic properties and functional-semantic case roles be unambiguously determined?

Case role detection will provide a way to identify individual semantic entities in a text. For instance, the sentence 'Invesco in merger talks with AIM Management' will be labelled with the following roles:

Invesco in merger	talks	with AIM Management
Sayer	*Verbal Process*	*Receiver*

Roles like the ones above constitute the semantic tagset. In the case of our algorithm, they are based on a simplified version of Halliday's systemic-functional grammar. The main advantage of our approach over previous, domain-specific

Overview

We consider case role detection to be a standard classification task. Each class is a concatenation of individual case roles into a meaningful pattern; the features that are used for training the classifier are lexical and morphosyntactic properties of the chunks corresponding to the case roles. Distinctive features that are

1 & 12

The training set and test set are derived from the new Reuters corpus. About 3000 files have been

2 & 11

In the preprocessing phases, the input text is cleaned, tagged and chunked with LTPOS. For each chunk, the

3

The training set is manually annotated, using an annotation tool specifically designed for assigning semantic roles to stretches of text. Iit took a job student about half a month to annotate 1500 examples. Our semantic tagset is currently based on a simplified version of Halliday's systemic-functional grammar, but will have to be replaced by a more

4

The annotated text is first saved into an intermediate XML structure, which still contains all information present in the annotated patterns. This will make it easier to replace the classifier and in a next phase of the project we might need the information that is lost in the conversion to the C4.5 input format for

5

The intermediate structure is converted to fixed-length feature vectors, taking lexical and morphosyntactic information about the lexical head of each phrase and of the boundaries of the left and right context of the head. Word order is encoded only implicitly. On the one hand, this feature selection will reduce the information to be processed by the classifier and rule out many irrelevant features; on the other hand, it

7

The data is processed by C4.5 and C4.5rules for each verb separately to discriminate features that are distinctive for individual case role patterns associated with the verb. The classification algorithm is likely to be replaced by a more appropriate one in a next phase of the project.

8

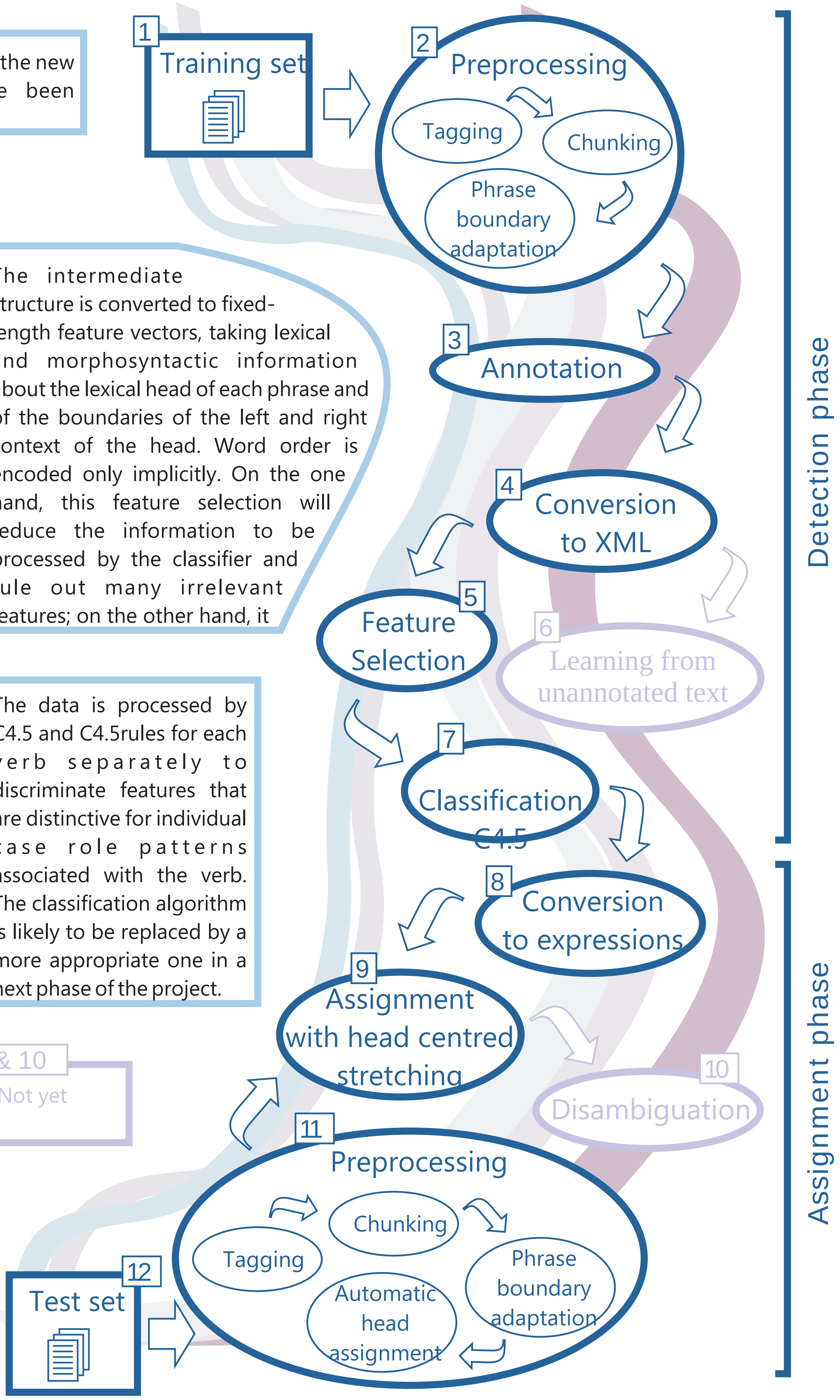
The rules resulting from the classification process are transformed into strings resembling regular expressions. This is necessary because of the variable length of the case role patterns and it partly compensates for

6 & 10

Not yet

9

The expressions are re-assigned to previously unseen text using head-centred stretching, i.e. the head of the case role pattern is assigned first and all peripheral constituents are matched from


1 Training set
2 Preprocessing
Tagging
Chunking
Phrase boundary adaptation
3 Annotation
4 Conversion to XML
5 Feature Selection
6 Learning from unannotated text
7 Classification C4.5
8 Conversion to expressions
9 Assignment with head centred stretching
10 Disambiguation
11 Preprocessing
Tagging
Chunking
Phrase boundary adaptation
Automatic head assignment
12 Test set
Detection phase
Assignment phase

Evaluation

Future improvements

The current setup is only a first try and therefore it should not be a surprise that initial results are far from impressive. However, many improvements are still possible:

- Replacing C4.5 by a more adequate classifier will boost results and is likely to reduce the annotation effort.
- Encoding relative word order as an explicit feature will make it easier to match corresponding case roles and will enable us to treat circumstances (which do not have a fixed position) more accurately.
- Using analytic tags instead of synthetic tags will improve rule generalization.
- Lexicons might also be employed to match partially correct patterns, to deal with idiomatic expressions and to use...
- Analogous reasoning to match partially correct patterns and to expand the pattern base.
- Designing a preference order for empty roles will match individual chunks correctly.
- Pattern selection, based on parameters such as pattern length,

We would like to thank IWT Flanders, Roularta Media Group, Kluwer Belgium, Land & Computing and ICMS for the research funding and