Acquisition of inflectional paradigms with minimal supervision

Radoslav Klíč

Seznam.cz a.s.

Outline

1. Introduction
2. Paramor
3. Seeding
4. Allomorphy
5. Results

Introduction

- The assignment: Acquisition of inflectional paradigms with minimal supervision. That may be useful if we have a plain-text corpus and no grammar-book of the language. Let's assume we can ask a native speaker to provide some examples of inflected words.
- The approach: Modification and extension of Paramor, an unsupervised paradigm learner.

Paradigms

- Classical Czech paradigms have slots for all combinations of relevant morphological categories.

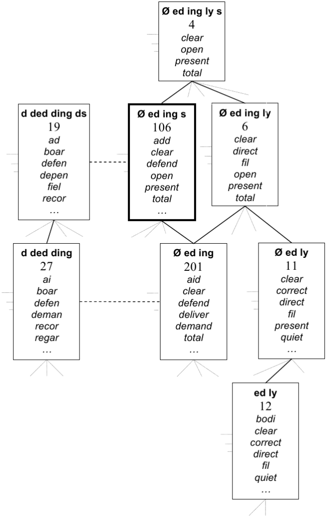
Case	Singular	Plural
nom	mat**k**+a	mat**k**+y
gen	mat**k**+y	mat**ek**+0
dat	mat**c**+e	mat**k**+ám
acc	mat**k**+u	mat**k**+y
voc	mat**k**+o	mat**k**+y
loc	mat**c**+e	mat**k**+ách
inst	mat**k**+ou	mat**k**+ami

- We don't know much about the grammar of the given language. Therefore we'll be happy with paradigms defined as a set of suffixes + set of stems e.g., (*a, y, e, u, o, ou, 0, ám, ách, ami*) + (*žen, matk, . . .*)

Paramor – Schemes

- In Paramor, partial paradigms are modelled by *schemes*.
- A scheme is defined by a set of its suffixes e.g., (*0, ed, ing, s*).
- The scheme's stem set is obtained deterministically by selecting all the candidate stems which form a word (present in the corpus) with all the schemes suffixes.
- Thus, adding a suffix can decrease the number of scheme's adherent stems (and cannot increase it). (More stems combine with (*0, ed, ing, s*) than with (*0, ed, ing, s, ly*))

Scheme lattice



Paramor algorithm

- Bottom-up search. Starts with single-suffix schemes and ascends the lattice. Stops when the c-stem ratio drops below 0.25.
- Scheme clustering. Similar schemes are joined into scheme clusters. Similarity is defined as similarity of produced <stem, suffix> pair sets. For example, schemes (*0, ly, ness*) and (*0, ly, er, est*) can be merged, as they share a lot of stem-suffix pairs like *deep* + *0*, *deep* + *ly*.
- Scheme cluster pruning.

Seeding

- I modified Paramor to be able to use manually entered input in the form of inflected word forms with marked morpheme boundary.
- Seed example: *matk/matc/matek + a, u, y / e / 0*
- Usage:
  - Add two-suffix schemes to the initial scheme set for bottom-up search. The suffix pairs are taken from the manual seed. (I use pairs because schemes with larger subsets need not be present in the corpus)
  - Protect some scheme clusters from discarding.
  - Induction of allomorphy rules.

Allomorphy

- Paramor does not recognise allomorphic stems. As a result, suffixes triggering phonological changes are often not selected in the bottom-up search, because they form words with different surface stems.
- For example, let's assume the bottom-up search on a Czech corpus reached a scheme *(a, y, u, ou)* with stems like *matk, noh* and tries to add *-e* suffix.
- In this case, stems where a phonological change is triggered (like *matk* → *matc*, *noh* → *noz*) will drop out after adding -*e*, which significantly decreases the c-stem ratio and causes the search to stop before adding -*e*.

Allomorphy – usage of the seed

- I induce rules from the manual seed which allow Paramor to join two or more surface stems into one.
- For example, from a seed entry

  $$\textit{politik/politic} + \textit{a, u, ovi, em, y, ů, ům} \;/\; \textit{i, ích}$$

  the following rule is generated:

  $$*k \leftrightarrow *c \;/\; \{\textit{a, u, ovi, em, y, ů, ům}\}, \{\textit{i, ích}\}$$

Evaluation

- The subjects of evaluation are clusters of words which are compared to lexemes in a lemmatised corpus. (Lexeme – set of all forms of one lemma.)
- The evaluation method is pair-wise. For each pair of words, I check whether they belong to the same lemma and whether they belong to the same cluster created by the algorithm. I count true/false positives and true/false negatives and from them I get precision and recall to compute the F-score.

Results

The F-score obtained with and without additional data:

Corpus	**no seed**	**seed**
cz	69.63	**72.99**
si	74.83	**75.61**
de	63.98	**64.52**
cat	62.74	**65.95**