An experiment compared the tagging of two languages: Czech, a highly inflected language with a high degree of ambiguity, and English. For Czech, the corpus was one gathered in the 1970s at the Czechoslovak Academy of Sciences; for English, it was the Wall Street Journal corpus. Results indicate 81.53 percent accuracy for Czech and 96.83 percent accuracy for English, representing a higher level of accuracy than expected for Czech. Several simple improvements in the Czech tagging system were identified.
A Simple Czech and English Probabilistic Tagger: a Comparison

Barbora Hladká, Jan Hajič

Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Praha 1
Tel.: +42 2 21914 288
Fax: +42 2 21914 309
E-mail: {hladka, hajic}@ufal.mff.cuni.cz
1. Introduction

Highly inflectional languages like Czech pose a special problem for morphology disambiguation (which is usually called tagging). For example, the ending -u is not only highly ambiguous, but at the same time it carries a complex information: it corresponds, e.g., to genitive singular for inanimate nouns, or dative singular for animate nouns, or accusative singular for feminine nouns, or first person singular present tense active participle for certain verbs.

Given the success of statistical methods in different areas including text tagging we wanted to try them even for the Czech language one of the main features of which is a rich inflection displaying a high degree of ambiguity. Originally we expected that the result would be plain negative, getting not more than about two thirds of the tags correct. However, as we show later, we got better results than we had expected.

We used the same statistical approach to tag both the English text and Czech text. For English, we obtained results comparable with the results presented in [Brill 1993] (who uses different methods). For Czech, we obtained results which are less satisfying than those for English results.

2. Data Used

2.1 For Czech

For training, we used the corpus collected at the beginning of the 70ies in the Czechoslovak Academy of Sciences. The corpus was originally hand-tagged, including the lemmatization and syntactic tags. The complete size of the corpus is 600k tokens. We had to do some cleaning and conversion, as we were interested in the words and tags only.

2.2 For English

For training, we used Wall Street Journal [Marcus, Santorini, Marcinkiewicz 1993]. We had to change the format of WSJ to prepare it for our tagging software.
3. Tags

3.1 Czech tags

The original tag system (in the hand-tagged corpus) was too detailed to use it directly. We disregarded all the other information (lemmatization and syntactic tags) from the training data. We used the traditional division into the part of speech tagger classes. Each class contains many tags for each combination of morphological categories. For a description of the tags for the part of speech classes see Table 1. The first letter represents the tag for the part of speech class and it is followed by the morphological categories for the given class. We used special tags for sentence boundaries, punctuation and "unknown tag". We used 1171 different tags in our experiment for Czech. They were manually derived from the training corpus.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>N</td>
<td>gender number case</td>
</tr>
<tr>
<td></td>
<td>Z</td>
<td>abbreviation</td>
</tr>
<tr>
<td>Adjectives</td>
<td>A</td>
<td>gender number case degree negation</td>
</tr>
<tr>
<td>Verbs</td>
<td>V</td>
<td>infinitive T negation</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>transgressive W number tense voice gender negation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>common</td>
</tr>
<tr>
<td></td>
<td></td>
<td>person number voice tense mood gender negation</td>
</tr>
<tr>
<td>Pronouns</td>
<td>P</td>
<td>personal person number case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 gender number case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gender number case</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>possessive gender-of-the-possessive number-of-the-possessive case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>person gender number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>svilj S gender number case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>se E case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>others D gender number case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>negation</td>
</tr>
<tr>
<td>Adverbs</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Conjunctions</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Prepositions</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Interjections</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Particles</td>
<td>K</td>
<td></td>
</tr>
</tbody>
</table>

Table 1
For example:
NMS1 (noun, masculinum animate, singular, nominative)
NNP7 (noun, neuter, plural, instrumental)
VTA (verb, infinitive, affirmative)
V3SAPOMA (verb, 3rd person, singular, active, present tense, indicative, mas. anim., affirmative)
PP2P7 (personal pronoun, 2nd person, plural, instrumental)
AFP32N (adjective, femin. plural, dative, comparative, negative)

3.2 English tags

We used The Penn Treebank tagset which contains 36 Part-Of-Speech tags and 12 other tags (for punctuation and the currency symbol). A detailed description is available in [Santorini 1990].

4. The algorithms

We have used Merialdo’s methods (described e.g., in [Merialdo 1992]). The tagging procedure selects a sequence of tags T for the sentence W:

$$\Phi : W \rightarrow T = \Phi(W).$$

In this case the optimal tagging procedure is

$$\Phi(W) = \operatorname{argmax}_{T} \Pr(T \mid W) = \operatorname{argmax}_{T} \Pr(T \mid W) \cdot \Pr(W) = \operatorname{argmax}_{T} \Pr(W \mid T) = \operatorname{argmax}_{T} \Pr(W \mid T) \cdot \Pr(T).$$

Our implementation is based on generating the (W, T) pairs by a probabilistic model using approximations of probability distributions Pr(W | T) and Pr(T).

The Pr(T) is based on tag bigrams, and Pr(W | T) is approximated as the product of Pr(wi | ti). The parameters have been estimated by the usual maximum likelihood training method, i.e. we approximated them as the relative frequencies found in the training data, smoothing them accordingly using the unigram frequencies and the uniform distribution.
5. The results

<table>
<thead>
<tr>
<th>corpus</th>
<th>Experiment for Czech</th>
<th>Experiment for English</th>
</tr>
</thead>
<tbody>
<tr>
<td>trainig data (tokens)</td>
<td>621 015</td>
<td>1 287 749</td>
</tr>
<tr>
<td>trainig data (words)</td>
<td>72 445</td>
<td>51 433</td>
</tr>
<tr>
<td>trainig data (tags)</td>
<td>1 171</td>
<td>45</td>
</tr>
<tr>
<td>training data (the</td>
<td>3,65</td>
<td>2,34</td>
</tr>
<tr>
<td>average number of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tags per token)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>test data (tokens)</td>
<td>1 294</td>
<td>1 294</td>
</tr>
<tr>
<td>incorrect tags</td>
<td>56</td>
<td>41</td>
</tr>
<tr>
<td>tagging accuracy</td>
<td>81,53%</td>
<td>96,83%</td>
</tr>
</tbody>
</table>

To illustrate the results of our tagging procedures, we present here an example from the tagged test text. The cases of incorrect tag assignment are denoted by boldface letters.

<table>
<thead>
<tr>
<th>tagged word</th>
<th>hand-assigned tag</th>
<th>result of the tagging programme</th>
</tr>
</thead>
</table>

**Czech test text**

jménem | Rjménem | NNS7
úv | NZ | NZ
Ksč | NZ | NZ
pozdravil | V3SAMOMA | NZ
Davisovou | NFS4 | NZ
Pavel | NMS1 | NMS1
Auersperg | NMS1 | NMS1
W_SB | T SB | T SB
účastníci | NMP1 | NMP1
shromáždění | NNS2 | NNS2

**English test text**

In | IN | IN
the | DT | DT
lengthy | JJ | JJ
discussion | NN | NN
that | IN | WDT
followed | VBD | VBD
, | , | ,
Mr. | NNP | NNP
Buffett | NNP | NNP
said | VBD | VBD
: | : | :
6. Conclusion

The results, however they might seem negative compared to English, are still better than our original expectations. We would like to improve current approach by another simple measures. For example, the average number of tags per token will increase after a morphological analyser is added as the front end to the tager (serving as the “supplier” of possible tags). We also plan to use trigrams instead of bigrams after we collect more data for Czech. Finally, certain tagset reductions be carried one, as the original tagset (even after the reductions mentioned above) is too detailed (in the sense that it distinguishes tags hardly distinguishable by human annotators). We are also working on independent predictions for certain grammatical categories and the lemma itself, but the final shape of the model has not yet been decided. This would mean to introduce constraints on possible combinations of morphological categories and take them into account when “assembling” the final tag.

References


I. DOCUMENT IDENTIFICATION:


Author(s): Heike Rettig (Ed.)

Corporate Source: 

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

For Level 1 Release:

- Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) and paper copy.

The sample sticker shown below will be affixed to all Level 1 documents.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

For Level 2 Release:

- Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but not in paper copy.

The sample sticker shown below will be affixed to all Level 2 documents.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: 

Organization/Address: Institut für deutsche Sprache 
R S 6-13 68161 Mannheim 
Postfach 101621 68016 Mannheim

Printed Name/Position/Title: Norbert Volz, M.A.

Telephone: 49 621 1581 437 
FAX: 49 621 1581 415

E-Mail Address: volz(at)ids-mannheim.de

Date: 28/11/97

(over)