The Context (not only) for Humans

Barbora Hladká

Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Prague
Czech Republic

e-mail: hladka@ufal.mff.cuni.cz

Abstract

Our context considerations will be practically oriented; we will explore the specification of a context scope in the Czech morphological tagging. We mean by morphological tagging/annotation the automatic/manual disambiguation of the output of morphological analysis. The Prague Dependency Treebank (PDT) serves as a source of annotated data. The main aim is to concentrate on the evaluation of the influence of the chosen context on the tagging accuracy.

1. English and Czech Tagging Experiments

The corpus-based approaches determine the amount of human work involved in the NLP tasks on the building of training data and on the coming up with an algorithm giving results as precise as possible. The ideas of context specification cannot be left out in the formulation of the algorithm. The scope of context must be specified according to the character of the particular NLP task. As we consider the nature of context from the perspective of the tagging application, an elementary unit we process is a word token. In general, there is no strict rule saying how many preceding and following word tokens we should look at to be sure that we tag the word token properly. Thus, let us have a look at the empirical experience.

STRATEGY	TAGGER ID	TRAINING DATA	TA (%)
Trigram Markov model (Merialdo, 1994)	MM_EN	Associated Press (955Kw)	97.0
Maximum Entropy (Ratnaparkhi, 1996)	ME_EN	WSJ (926Kw)	96.6
Exponential model (Hajič, Hladk´a, 1998b)	EXP_EN	WSJ (1.2 Mw)	96.8
Memory-based (Daelemans, Zavrel, 1996)	MB_EN	WSJ (2Mw)	96.4
Rule-based (Brill, 1993b)	RB_EN	WSJ (600Kw)	96.9
Neural Networks (Schmid, 1994)	NE_EN	WSJ (2Mw)	96.2

Table 1: Tagging experiments on English

STRATEGY	TAGGER ID	TRAINING DATA	TA (%)
Bigram Markov model	MM_CZ$_{bi}$	PDT (300Kw)	92.50
Trigram Markov model	MM_CZ$_{bi}$	PDT (300Kw)	93.38
Exponential model	EXP_CZ	PDT (300Kw)	93.85

Table 2: Tagging experiments on Czech

Let $W = w_1 w_2 w_3 ... w_n$ be an input text to be tagged. As all the presented tagging strategies tag the input text in the left-to-right direction, a word token w_i is processed when the word tokens $w_1 w_2 ... w_{i-1}$ have already been tagged - $w_1|t_1 ... w_{i-1}|t_{i-1} w_i w_{i+1} ... w_n$[1] For the currently processed word token w_i, the context $c(w_i)$ of the representative corpus-based tagging strategies for tagging Czech (MM_CZ$_{tri}$, MM_CZ$_{bi}$, EXP_CZ taggers - see Tab. 1[2]) and English (MM_EN, ME_EN, EXP_EN, MB_EN, RB_EN, NE_EN taggers - see Tab. 2) can be expressed as follows:

MM_EN - $c(w_i) = \{w_{i-2}, t_{i-2}, w_{i-1}, t_{i-1}\}$

MM_CZ$_{tri}$ - $c(w_i) = \{t_{i-2}, t_{i-1}\}$

MM_CZ$_{bi}$ - $c(w_i) = \{t_{i-1}\}$

ME_EN - $c(w_i) = \{w_{i-2}, t_{i-2}, w_{i-1}, t_{i-1}, w_{i+1}, w_{i+2}\}$

EXP_CZ, EXP_EN - $c(w_i) = \{w_{i-4}, ma_{i-4}, w_{i-3}, ma_{i-3}, w_{i-2}, ma_{i-2}, w_{i-1}, ma_{i-1}, w_{i+1}, ma_{i+1}, w_{i+2}, ma_{i+2}, w_{i+3}, ma_{i+3}, w_{i+4}, ma_{i+4}\}$

MB_EN - not directly specified

RB_EN - $c(w_i) = \{w_{i-3}, t_{i-3}, w_{i-2}, t_{i-2}, w_{i-1}, t_{i-1}, w_{i+1}, w_{i+2}, w_{i+3}\}$

NE_EN - $c(w_i) = \{w_{i-3}, t_{i-3}, w_{i-2}, t_{i-2}, w_{i-1}, t_{i-1}, w_{i+1}, w_{i+2}\}$

Observing the given description, the Markov models are locally (processing w_i) based only upon the left-hand

[1] As the EXP tagger operates on a subtag level let ma$_i$ consist of all possible values of *i*-th morphological category within the positional tag.

[2] (Hladk´a, 2000) provides very detailed view on the issue of Czech language tagging.

side context[3] while the other strategies look not only at the left positions but consider also the right-hand side context. Some authors offer practical experience with a modification of the context scope. In the paper (Schmid, 1994), the author describes the context shrinking to two preceding and one following words together with their tags which causes accuracy reduction only by 0.1%. Enlarging the context gave no improvement. The authors (Daelemans, Zavrel, 1996) do not specify directly the context scope, but they construct a distance metrics between similar environments within modest contexts. We can conclude that the enumerated contexts as a whole are limited up to 4 positions to the left/right.

2. The Context for Humans

At the starting point of the tagging procedure, all tagging strategies are given the same input text. The input text (as a whole) is understood as a whole text context. Consequently, the tagging strategies select from the whole text context any subcontext over which they process the given word token. Let us limit the subcontext to the word tokens $(w_1, w_2, ..., w_{i-1})$ preceding the currently processed word token (w_i) within the input text. For a vocabulary size *n* there are n^{i-1} different subcontexts (for ex. *n = 1,000* and *i = 4* then $n^3 = 10^9$). The problem which immediately appears concerns the matrices (of n^{i-1} order) representing the counts of particular subcontexts within the training corpus. With regard to the astronomically large number of such subcontexts, a vast majority of the possible subcontexts will never occur in Czech (or other natural language) and that is the reason why the given matrices are sparse. Nevertheless, the computational linguists' effort is directed to deal with sparseness of data being connected with context specification.

None of the representative corpus-based tagging methods do achieve the magic point of 100% performance (Table 1, 2). It is supposed that the context can reveal almost all the secrets of a language. We stress *almost*, in some cases the context is not enough to specify the function/meaning of a word form. As we are interested in context-based models of a language, the magic point of such models cannot be 100% because the world knowledge which is hidden somewhere between the lines cannot be read in the set of word forms and tags.

The bigram and trigram MMs employ the smallest left-hand side context size relatively to the other corpus-based methods; at the same time, their performances are the best (Table 1, 2). We believe that a further improvement of MMs lies in a better selection of the analysed context. Not to limit ourselves only to experiments modifying the context size and in order to discover certain guidelines we explore how people handle the information coming from the predefined left-hand side context.

2.1. Prerequisites

The annotation of the test file was assigned to a group of 5 students: 2 undergraduate students (S1, S2) with rich experience learned during the annotation of the PDT; 3 computational linguistics graduate students (S3, S4, S5) - one of them (S5) with an experience with various tagging strategies and one of them (S4) being bilingual not educated in Czech. The test file that was given to the students comprised a 283 word token subset (141 unambiguous tokens and 142 ambiguous tokens) of the test file which we used in the tagging experiments mentioned above (MM_CZ$_{tri}$, MM_CZ$_{bi}$, EXP_CZ). For purposes of evaluation of the tagging and annotation, the given test file was annotated independently by another annotator upon an unlimited context.

2.1.1. Formalism

Let S = $w_1 w_2 ... w_s$ be a sentence[4](a sequence of word tokens) we tag/annotate in the left-to-right direction, $S_{tokens} = (w_i)_{i=1..s}$ be a list of word tokens occurring in the sentence S. While tagging the *i*-th word, the *i-1* preceding word tokens are already tagged by tags $t_1, t_2, ..., t_{i-1}$; let T be a list of tags $(t_j)_{j=1...i-1}$.

The contexts which come into play during the experiments of annotation (BC, TC, SC) and the experiments of tagging (TTC, BTC) can be defined as functions:

- **Bigram Context (BC)** as a function

  $BC : S_{tokens} \rightarrow 2^{S_{tokens}}$,
  $BC(w_i) = \{w_{i-1}\}, BC(w_1) = \emptyset$

- **Tag Bigram Context (TBC)** as a function

  $TBC : S_{tokens} \rightarrow 2^T$,
  $TBC(w_i) = \{t_{i-1}\}, TBC(w_1) = \emptyset$

- **Trigram Context (TC)** as a function

  $TC : S_{tokens} \rightarrow 2^{S_{tokens}}$,
  $TC(w_i) = \{w_{i-1}, w_{i-2}\}, TC(w_1) = \emptyset$,
  $TC(w_2) = \{w_1\}$

- **Tag Trigram Context (TTC)** as a function

  $TTC : S_{tokens} \rightarrow 2^T$,
  $TTC(w_i) = \{t_{i-1}, t_{i-2}\}, TTC(w_1) = \emptyset$,
  $TTC(w_2) = \{t_1\}$

- **Sentence Context (SC)** as a function

  $SC : S_{tokens} \rightarrow 2^{S_{tokens}}$,
  $SC(w_i) = \{w_1, ..., w_{i-1}\}, SC(w_1) = \emptyset$

To illustrate the defined terms, let us assume a sample of the sentence fragment *O další Stříbrné medvědy se podělily ...* [lit. about – further – Silver – Bears – Refl. – they-shared ..., E. The remaining (Prizes of) Silver Bears were obtained by ...] and let us suppose that the first four word tokens are already tagged O|RR--4---------- další|AAMP4----1A---- Stříbrné|AAMP4----1A----

[3] In the end, the incorporation of the Viterbi algorithm to find the best tag sequence means the usage of the right-hand side context.

[4] We consider a context within a sentence, we do not cross the sentence boundaries.

medvědy|NNMP4-----A----. Then the word token *se* is to be tagged/annotated. According to the chosen particular context, the word token *se* is being processed within the context information embodied in one of the sets BC(se) = {medvědy}, TC(se) = {Stříbrné, medvědy}, SC(se) = {O, další, Stříbrné, medvědy}, TTC(se) = {AAMP4----1A----, NNMP4-----A----}, BTC(se) = {NNMP4-----A----}.

2.2. How Humans Treat the Context Information

A specially developed tool for morphological annotation (Hajič, Hladká, 1998b), which offers an easy disambiguation of lemmas and tags (which are outputs of the automatic morphological analysis), was used as a disambiguation tool, which displays, for the currently annotated ambiguous word token, its morphological information and the whole text context. For our aims, we have modified the disambiguation tool in the sense of the visibility of a partial context; in case of Bigram Context only the previous word token is visible, for Trigram Context only two previous word tokens are, and finally, for Sentence Context the preceding word tokens up to the beginning of the sentence are at the annotator's disposal. We have to stress that unambiguous word tokens remain obviously untouched by the annotator and while annotating the given ambiguous word token the annotators have no information on the assigned tags of the word tokens which are included in the context; annotators just suggest a hypothesis of the tags of the context word tokens themselves. On the other hand, the presented Markov models working over Tag Trigram/Bigram Context do not deal with the word tokens.

3. Discussion of the Results

Table 3 provides information on the evaluation of the annotation and tagging of the given test file. Reading the table horizontally, we observe that all the students are getting better as the context enlarges. Reading the table vertically, we speculate that the learned experience in the course of the annotation over the whole context comes into play (students S1, S2 vs. students S3, S4, S5). On the other hand, the knowledge of the tagging methods seems not to be so important (student S5). The bigram MMs beat the students annotating over the bigram context TBC. However, the situation is inverse for the trigram contexts TTC, TC - annotation almost beats tagging.

Tables 4 and 5 give a detailed view on the annotation/tagging on a subtag level[5]. A more interesting observation concerns the way how the error rate over these MCs changes as the context enlarges.

context	BC	TC	SC	TTC	TBC
annotator/ tagger	# of incorrectly processed ambiguous word tokens out of 142 ambiguous				
S3	36	20	16	-	-
S4	47	32	27	-	-
S1	26	20	9	-	-
S2	16	13	7	-	-
S5	29	20	17	-	-
MM_CZ$_{tri}$	-	-	-	20	-
MM_CZ$_{bi}$	-	-	-	-	24

Table 3: The evaluation of tagging and annotation over the predefined contexts

annotator/ tagger	context	POS	SubPOS
S3	BC	0.71	0.71
	TC	0.35	0.35
	SC	0.00	0.35
S4	BC	1.06	1.41
	TC	1.77	2.12
	SC	0.35	0.71
S1	BC	0.35	0.71
	TC	0.35	0.71
	SC	0.00	0.35
S2	BC	0.35	0.35
	TC	0.35	0.35
	SC	0.35	0.35
S5	BC	0.06	1.77
	TC	0.00	0.35
	SC	0.35	0.71
MM_CZ$_{bi}$	BTC	0.71	0.71
MM_CZ$_{tri}$	TTC	0.71	0.71

Table 4: Error rates (%) over the POS, SubPOS

Looking at Tab. 6, the numbers represent decreasing/increasing (positive/negative numbers) of the error rates over the MCs gender, number, case for each student and the MM taggers. For example, for student S3, the error rate over gender decreases by 0.36% if the bigram context (BC) is enlarged to the trigram context (TC) and at the same time it decreases by 0.7% if the trigram context (TC) is enlarged to the sentence context (SC). Given the Czech typical word order and given the assumed left-hand side contexts, the improvement of the case error rate is more expressive than the changes of the gender and number error rates. Again, given the Czech word order, it is necessary to include the right-hand side context to identify the gender and number of the word token.

The strategy of human annotation described above can be understood only as a simulation of the MMs. The humans work with the left-hand side context from the beginning till the end; the MMs assign to the currently processed word token tags with regard to the left-hand side context as well, but the incorporation of the Viterbi algorithm to find the best tag sequence which means, in fact, the usage of the right-hand side context in fact.

Putting together this fact and the insufficiently represen-

[5] We present only the most problematic morphological categories (MCs) - gender, number, case - together with part of speech (POS) and sub-part of speech (SUBPOS).

annotator/ tagger	context	g	n	c
S3	BC	4.95	3.18	8.13
	TC	4.59	1.77	2.83
	SC	3.89	1.41	2.47
S4	BC	6.36	3.53	13.43
	TC	4.24	3.18	8.83
	SC	4.95	2.47	6.36
S1	BC	4.59	1.77	5.65
	TC	2.83	1.77	4.24
	SC	2.12	1.06	1.41
S2	BC	2.83	0.35	3.53
	TC	1.06	0.35	3.53
	SC	1.41	0.35	1.06
S5	BC	6.36	2.47	6.01
	TC	4.95	2.12	3.89
	SC	4.59	2.47	3.89
MM_CZ$_{bi}$	BTC	2.47	0.71	6.71
MM_CZ$_{tri}$	TTC	2.12	0.35	5.30

Table 5: Error rates (%) over the `gender, number, case`

morphological category		g	n	c
annotator/ tagger	context enlarging	the error rate improvement (%)		
S3	TC←BC	0.36	1.41	5.3
	SC←TC	0.7	0.36	0.36
S4	TC←BC	-0.71	0.35	4.6
	SC←TC	1.41	0.71	2.47
S1	TC←BC	1.76	0.00	1.41
	SC←TC	0.71	0.71	2.83
S2	TC←BC	1.77	0.00	0.00
	SC←TC	-0.35	0.00	2.47
S5	TC←BC	1.68	0.35	2.12
	SC←TC	0.36	-0.35	0.00
MM_CZ	TTC←TBC	0.35	0.36	1.41

Table 6: The error rate changes (%) due to the context enlarging

tative size of the test sample we cannot make any definite conclusions. On the other hand, the presented results offer the idea that the sentence context (SC) can be sufficient for successful context-based approaches. We speculate that it is not necessary to take the sentence context (SC) as a whole, but dynamically to select a trigram subcontext from the sentence context. The next step toward the specification of dynamic selection strategy will concern the type of information were used in human deciding limited by the SC (like for the human improvement of the speech recognizer's output, see (Brill et al., 1998).

4. Acknowledgement

The results described herein have been obtained partially within the individual grant of the OSF/HESP No. 135/1998 and partially within the Ministry of Education Project No. VS9615.

5. References

E. Brill. Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach. In *Proceedings of the 3rd International Workshop on Parsing Technologies*, Tilburg, Netherlands, 1993.

E. Brill, R. Florian, J. C. Henderson and L. Mangu. Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling? In *Proceedings of the AC/COLING Conference*, pp. 186-190, Montreal, Canada, 1998.

W. Daelemans and J. Zavrel. MBT: A Memory-Based Part of Speech Tagger-Generator. In *Proceedings of the Workshop on Very Large Corpora*, pp. 14-27, Copenhagen, Denmark, 1996.

J. Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Charles University Press - Karolinum, in press.

J. Hajič and B. Hladká. Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 111-118, Washington, USA, 1997.

J. Hajič and B. Hladká. Czech Language Processing - PoS Tagging. In *Proceedings of the First International Conference on Language Resources & Evaluation*, pp. 931-936, Granada, Spain, 1998.

J. Hajič and B. Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pp. 483-490, Montreal, Canada, 1998.

B. Hladká. *Czech Language Tagging*. PhD Thesis at the Institute of Formal and Applied Linguistics, Charles University, prague, Czech Republic, 2000.

B. Merialdo. Tagging English Text with a Probabilistic Model. In *Computational Linguistics*, 20(2), pp. 155-171, 1994.

A. Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the First Empirical Methods in Natural Language Processing Conference*, pp. 133-141, Philadelphia, USA, 1996.

H. Schmid. Part-Of-Speech Tagging with Neural Networks. In *Procedings of the 15th COLING Conference*, pp. 172-176, Kyoto, Japan, 1994.