Reference Data for Czech Collocation Extraction

Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

We introduce three reference data sets provided for the MWE 2008 evaluation campaign focused on ranking MWE candidates. The data sets comprise bigrams extracted from the *Prague Dependency Treebank* and the *Czech National Corpus*. The extracted bigrams are annotated as collocational and non-collocational and provided with corpus frequency information.

1. Motivation

Gold standard reference data is absolutely essential for empirical evaluation. For many tasks of Computational Linguistics and Natural Language Processing (such as machine translation or word sense disambiguation) standard and well designed reference data sets are widely available for evaluation and development purposes. Since this has not been the case for the task of collocation extraction, we decided to develop a complete test bed on our own with the aim to use it for evaluation of methods for collocation extraction (Pecina and Schlesinger, 2006).

In this paper we presents three sets of bigrams extracted from the *Prague Dependency Treebank*: one set consists of dependency (syntactical) bigrams, the second one of surface (adjacent) bigrams, and the third one contains instances of the second one in the *Czech National Corpus*. The extracted bigrams are annotated as collocational and non-collocational (and also assigned to finer-grained categories). The reference sets are associated with corpus frequency information for easy computation of association measure scores. All the data sets are publicly available from the MWE wiki page[1].

2. Prague Dependency Treebank

The *Prague Dependency Treebank 2.0* (PDT) is a moderate sized corpus provided with manual morphological and syntactic annotation. By focusing only on two-word collocations, PDT provides sufficient evidence of observations for a sound evaluation. By default the data is divided into training, development, and evaluation sets. We ignored this split and used all data annotated on the morphological and analytical layer: a total of 1 504 847 tokens in 87 980 sentences and 5 338 documents.

2.1. Treebank Details

The Prague Dependency Treebank[2] has been developed by the Institute of Formal and Applied Linguistics and the Center for Computational Linguistics, Charles University, Prague and it is available from LDC[3] (catalog number LDC2006T01). It contains a large amount of Czech texts with complex and interlinked annotation on morphological, analytical (surface syntax), and tectogrammatical (deep syntax) layer. The annotation is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs.

Morphological Layer

On the morphological layer each word form (token) is assigned a *lemma* and a *morphological tag*. Combination of the lemma and the tag uniquely identifies the word form. Two different word forms differ either in lemmas or in morphological tags. Lemma has two parts. First part, the *lemma proper*, is a unique identifier of the lexical item. Usually it is the base form (e.g. first case singular for a noun, infinitive for a verb, etc.) of the word, possibly followed by a number distinguishing different lemmas with the same base forms (different word senses). Second part is optional. It contains additional information about the lemma (e.g. semantic or derivational information). Morphological tag is a string of 15 characters where every position encodes one morphological category using one character. Description of the categories and range of their possible values are summarized in Table 1. Details of morphological annotation can be found in (Zeman et al., 2005).

Pos	*Name*	*Description*	*# Values*
1	**POS**	**Part of speech**	**12**
2	SubPOS	Detailed part of speech	60
3	**Gender**	**Gender**	**9**
4	Number	Number	5
5	Case	Case	8
6	PossGender	Possessor's gender	4
7	PossNumber	Possessor's number	3
8	Person	Person	4
9	Tense	Tense	5
10	**Grade**	**Degree of comparison**	3
11	**Negation**	**Negation**	2
12	Voice	Voice	2
13	Reserve1, 2	Reserve	-
14	Reserve2	Reserve	-
15	Var	Variant, style	10

Table 1: Morphological categories encoded in Czech tags.

Analytical Layer

Analytical layer of PDT serves to encode sentence *dependency structures*. Each word is linked to its *head word* and assigned its *analytical function* (dependency type). If we think of a sentence as a graph with words as nodes and dependency relation as edges, the dependency structure is

[1] http://multiword.wiki.sourceforge.net/
[2] http://ufal.mff.cuni.cz/pdt2.0/
[3] http://www.ldc.upenn.edu/

Id	*Form*	*Lemma*	*Full Tag*	*Parent Id*	*Afun*	*Id*	*Lemma Proper*	*Reduced Tag*	*Parent Id*	*Afun*
1	Zbraně	zbraň	NNFP1-----A----	0	ExD	1	zbraň	NF-A	0	Head
2	hromadného	hromadný	AANS2----1A----	3	Atr	2	hromadný	AN1A	3	Atr
3	ničení	ničení_^(*3it)	NNNS2-----A----	1	Atr	3	ničení	NN-A	1	Atr

Table 2: Example of annotated and normalized expression (*weapons of mass destruction*). A normalized form consists of a lemma proper (lemma without technical suffixes) and a reduced morphological tag (positions 1, 3, 10, 11 of the full tag).

a tree – a directed acyclic graph having one root. Details of analytical annotation can be found in (Hajič et al., 1997).

2.2. Collocation Candidate Data Sets

Two collocation candidate data sets were obtained from PDT. Both were extracted from morphologically normalized texts and filtered by a frequency filter and a part-of-speech filter. Details of these steps are the following:

Morphological Normalization
The usual role of morphological normalization is to canonize morphological variants of words so that each word (lexical item) can be identified regardless its actual morphological form. This technique has been found very beneficial for example in information retrieval, especially on morphologically rich languages such as Czech.Two basic approaches to this problem are: *stemming*, where a word is transformed (usually heuristically) into its *stem* which often does not represent a meaningful word, and *lemmatization*, where a word is properly transformed into its base form (lemma) by means of morphological analysis and disambiguation.
The latter approach seems more reasonable in our case (manually assigned lemmas are available in PDT) but it is not completely adequate. By transforming words only into lemmas we would loose some important information about their lexical senses that we want to preserve and use to distinguish between occurences of different collocation candidates. For example *negation* and *grade* (degree of comparison) significantly change word meanings and differentiate between collocation candidates (eg. “secure area” vs. “insecure area”, "big mountain“ vs. ”(the) highest mountain“). Indication of such morphological categories is not encoded in a lemma but rather in a tag. With respect to our task, we decided to normalize word forms by transforming them into combination of a *lemma* (lemma proper, in fact; the technical suffixes in PDT lemmas are omitted) and a *reduced tag* that comprises the following morphological categories: *part-of-speech*, *gender*, *grade*, and *negation* (highlighted in Table 1). For similar reasons and also in order to decrease granularity of collocation candidates, we simplified the system of Czech analytical functions by merging some of them into one value.

Part-of-Speech Filtering
A part-of-speech filter is a simple heuristic that improves results of collocation extraction methods a lot (Justeson and Katz, 1995): the collocation candidates are passed through a filter which only lets through those patterns that are likely to be 'phrases' (potential collocations). Justenson and Katz (1995) filtered the data in order to keep those that are more likely to be collocations than others; for bigram collocation extraction they suggest to use only patterns A:N (adjective–noun) and N:N (noun–noun). We, however, deal with a broader notion of collocation in our evaluation and this constraint would be too limitative. We filter out candidates having such part-of-speech patterns that *never* form a collocation (at least in our data), in other words to keep the cases with part-of-speech patterns that can *possibly* form a collocation. This step does not effect the evaluation because it can be done prior to all extraction methods. The list of employed patterns is presented in Table 3. It was proposed congruently by our annotators before the annotation process described in Section 2.3.

Frequency Filtering
As mentioned earlier our motivation to create the reference data set was empirical evaluation of methods for collocation extraction. To ensure that the evaluation is not biased by low-frequency data, we limit ourselves only on collocation candidates occurring in PDT more than five times. The less frequent candidates do not meet the requirement of sufficient evidence of observations needed by some methods (they assume normal distribution of observations and/or become unreliable when dealing with rare events). Moore (2004) argues that these cases comprise majority of all the data (the well-known Zipfian phenomenon) and should not be excluded from real-world applications.

PDT-Dep
Dependency trees from the treebank were broken down into the dependency bigrams. From all PDT sentences we obtained a total of 635 952 different dependency bigram types (494 499 of them were singletons). Only 26 450 of them occur in the data more than five times. After applying the frequency and part-of-speech pattern filter we obtained a list of 12 232 collocation candidates (consisting of a normalized head word and its modifier, plus their dependency type) further referred to as *PDT-Dep*.

PDT-Surf
Although collocations form syntactic units by definition, we can attempt to extract collocations also as *surface bigrams* (pairs of adjacent words) without guarantee that they form such units but with the assumption that majority of bigram collocations can not be modified by insertion of another word and in text they occur as surface bigrams (Manning and Schütze, 1999, chapter 5). This approach does not require the source corpus to be parsed, which is usually a time-consuming process accurate only to a certain extent. A total of 638 030 surface bigram types was extracted from PDT, 29 035 of them occurred more then five times and after applying the part-of-speech filter we obtained a list of 10 021 collocation candidates (consisting of normalized components) further referred to as *PDT-Surf*. 974 of these bigrams do not appear in *PDT-Dep* test sets (if

we ignore the syntactical information).

2.3. Manual Annotation

Three educated linguists, familiar with the phenomenon of collocations, were hired to annotate the reference data sets extracted from PDT in paralell. To consolidate their notion of collocation we adopt the definition from Choueka (1988): "A collocation expression is a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." It is relatively wide and covers a broad range of lexical phenomena such as idioms, phrasal verbs, light verb compounds, technological expressions, proper names, and stock phrases. It requires collocations to be syntactic units – subtrees of sentence dependecy trees in case of dependecy syntax used in PDT.

The dependency bigrams in *PDT-Dep* were assessed first. The annotation was performed independently and without knowledge of context. To minimize the cost of the process each collocation candidate was presented to each annotator only once although it could appear in many different contexts. The annotators were instructed to judge any bigram which could *eventually* appear in context where it has a character of collocation, as a *collocation*. E.g. idiomatic expressions were judged as collocations although they can also occur in contexts where they have a literal meaning. Similarly for other types of collocations. As a result the annotators were quite liberal in their judgments which we exploited in combining their outcomes.

During the assessment the annotators also attempted to classify each collocation into one of the following categories. This classification, however, was not intended as a result of the process but rather as a way how to clarify and simplify the annotation. Any bigram that can be assigned to any of the categories was considered a collocation.

1. stock phrases
   zásadní problém (major problem), konec roku (end of a year)
2. names of persons, organizations, geographical locations, and other entities
   Pražský hrad (Prague Castle), Červený kříž (Red Cross)
3. support verb constructions
   mít pravdu (to be right), činit rozhodnutí (make decision)
4. technical terms
   předseda vlády (prime minister), očitý svědek (eye witness)
5. idiomatic expressions
   studená válka (cold war), visí otazník (hanging question mark ~ open question)

The surface bigrams from *PDT-Surf* were annotated in the same fashion but only those collocation candidates that do not appear in *PDT-Dep* were actually judged (974 items). Technically we removed the syntactic information from *PDT-Dep* data and transfer the annotations to *PDT-Surf*, if a surface bigram from *PDT-Surf* appears also in *PDT-Dep* it is assigned the same annotation from all three annotators.

Inter-annotator Agreement

The interannotator agreement among all the categories of collocations (plus a 0 category for non-collocations) was relatively low: the average accuracy between two annotators on *PDT-Dep* was as low as 72.88%, the average Cohen's κ was estimated as 0.49. This demonstrates that the notion of collocation is very subjective, domain-specific, and also somewhat vague. Since we did not distinguish between different collocation categories – ignoring them (considering only two categories: *true collocations* and *false collocations*) increased the average accuracy up to 80.10% and the everage Cohen's κ to 0.56. The three annotators were employed to get a more precise and objective idea about what can be considered a collocation by combining their independent outcomes. Only those candidates that *all* three annotators recognized as collocations (of any type) were considered *true collocations* (full agreement required). The *PDT-Dep* reference data set contained 2 557 such bigrams (21.02%) and *PDT-Surf* data set 2 293 (22.88%). For comparison of these reference data set see Figure 1.

Pattern	*Example*	*Translation*
A:N	trestný čin	*criminal act*
N:N	doba splatnosti	*term of expiration*
V:N	kroutit hlavou	*shake head*
R:N	bez problémů	*no problem*
C:N	první republika	*First Republic*
N:V	zranění podlehnout	*succumb*
N:C	Charta 77	*Charta 77*
D:A	volně směnitelný	*free convertible*
N:A	metr čtvereční	*squared meter*
D:V	těžce zranit	*badly hurt*
N:T	play off	*play-off*
N:D	MF Dnes	*MF Dnes*
D:D	jak jinak	*how else*

Table 3: Part-of-speech patterns for filterring collocation candidates (A – adjectives, N – nouns, C – numerals, V – verbs, D – adverbs, R – prepositions, T – particles).

3. Czech National Corpus

At the time of multi-billion word corpora, a corpus of the size of PDT is certainly not sufficient for real-world applications. We attempted to extract collocations also from larger data – a set of 242 million tokens from the *Czech National Corpus*. This data, however, lacks of any manual annotation, hence we settle for an automatic part-of-speech tagging (Hajič, 2004) and extracted collocation candidates as surface bigrams similarly as in the case of *PDT-Surf*.

3.1. Corpus Details

The *Czech National Corpus* (CNC) is an academic project with the aim to build up a large computer-based corpus, containing mainly written Czech[4]. The data we used comprises of two synchronous (containing contemporary written language) corpora SYN2000 and SYN2005 (ICNC, 2005) each containing about 100 million running words (excluding punctuation).

3.2. Automatic Preprocessing

SYN2000 and SYN2005 are not manually annotated, neither on morphological nor analytical layer. Manual anota-

[4]http://ucnk.ff.cuni.cz/

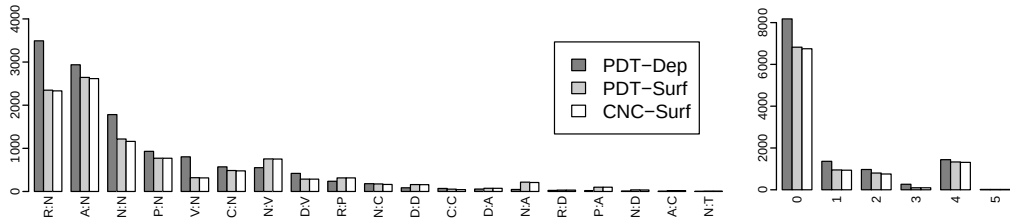

Figure 1: Part-of-speech pattern distribution in the reference data sets(left) and distribution of collocation categories in the reference data sets assigned by one of the annotators (right).

tion of such amount of data would be unfeasible. These corpora, however, are processed by a part-of-speech tagger.

3.3. Collocation Candidate Data Set

CNC-Surf

From the total of 242 million tokens from SYN2000 and SYN2005 we extracted more then 30 million surface bigrams (types). We followed the same procedure as for PDT reference data and after applying the part-of-speech and frequency filters, the list of collocation candidates contained 1 503 072 surface bigrams. Manual annotation of such amount of data was infeasible. To minimize the cost we selected only a small sample of it – already annotated bigrams from the *PDT-Surf* reference data set – a total of 9 868 surface bigrams further called *CNC-Surf*. All these bigrams appear also in *PDT-Surf*, the remaining 153 do not occur in the corpora more than five times. The major difference is only in the frequency counts provided with the data set. This reference data set contains 2 263 (22.66%) *true collocations* – candidates that all three annotators recognized as collocations (of any type). For comparison with the reference data sets extracted from PDT see Figure 1.

Reference Data Set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*
sentences	87 980	87 980	15 934 590
tokens	1 504 847	1 504 847	242 272 798
words (no punctuation)	1 282 536	1 282 536	200 498 152
bigram types	635 952	638 030	30 608 916
after frequency filtering	26 450	29 035	2 941 414
after part-of-speech filtering	12 232	10 021	1 503 072
collocation candidates	12 232	10 021	9 868
sample size (%)	100	100	0.66
true collocations	2 557	2 293	2 263
baseline precision (%)	21.02	22.88	22.66

Table 4: Statistics of the three reference data sets and the corpora they were extracted from.

4. Summary

We prepared three reference data sets for the task of identifying collocation candidates. All of them consist of two-word collocation candidates. *PDT-Dep* and *PDT-Surf* were extracted from the manually annotated Czech *Prague Dependency Treebank* and differ only in the character of bigrams. *PDT-Dep* consists of dependency bigrams and *PDT-Surf* of surface bigrams. Both were filtered by the same part-of-speech pattern filter and frequency filter. Manual annotation was done exhaustively – no sampling was needed, true collocations are indicated in all data.

CNC-Surf reference data set was extracted from much larger data from the *Czech National Corpurs* and comprises surface bigrams also appearing in *PDT-Surf*. It can be considered as a random sample from the full set of collocation candidates filtered by the same part-of-speech pattern filter and frequency filter as the PDT reference data.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838.

5. References

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. 1997. A manual for analytic layer tagging of the prague dependency treebank. Technical Report TR-1997-03, ÚFAL MFF UK, Prague, Czech Republic.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, volume 1. Charles University Press, Prague.

ICNC. 2005. Czech national corpus. Institute of the Czech National Corpus Faculty of Arts, Charles University, Praha, http://ucnk.ff.cuni.cz.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on EMNLP*, Barcelona, Spain.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia.

Dan Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Emil Jeřábek, and Barbora Vidová Hladká. 2005. A manual for morphological annotation, 2nd edition. ufal technical report no. tr-2005-27. Technical Report TR-2005-27, ÚFAL MFF UK, Prague, Czech Republic.