Comparison of Different Lemmatization Approaches through the Means of Information Retrieval Performance*

Jakub Kanis[1] and Lucie Skorkovská[1]

Univ. of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic
{jkanis,lskorkov}@kky.zcu.cz

Abstract. This paper presents a quantitative performance analysis of two different approaches to the lemmatization of the Czech text data. The first one is based on manually prepared dictionary of lemmas and set of derivation rules while the second one is based on automatic inference of the dictionary and the rules from training data. The comparison is done by evaluating the mean Generalized Average Precision (mGAP) measure of the lemmatized documents and search queries in the set of information retrieval (IR) experiments. Such method is suitable for efficient and rather reliable comparison of the lemmatization performance since a correct lemmatization has proven to be crucial for IR effectiveness in highly inflected languages. Moreover, the proposed indirect comparison of the lemmatizers circumvents the need for manually lemmatized test data which are hard to obtain and also face the problem of incompatible sets of lemmas across different systems.

1 Introduction

The task of automatic lemmatization, i.e. finding the "lexical headword" of a given word form, is one of the tasks that are especially important for the highly inflected languages such as Czech where the abundance of word forms pertaining to a single lemma complicates many of natural language processing tasks, ranging from the language modeling (where it causes unfavorable fragmentation of the training data) to the tasks of keyword spotting and information retrieval (IR), where the (very frequent) mismatch between the word form used in the query and the word forms occurring in the searched collection prevents many keyword occurrences and/or relevant documents from being found. The importance of the lemmatizers for IR effectiveness that was revealed by the previous experiments together with the fact that the intrinsic evaluation of the lemmatizers (i.e. measuring their performance on the manually annotated gold standard

* This research was supported by the Min. of Education of the Czech Republic, project. No. MŠMT LC536, by the grant of the University of West Bohemia, project No. SGS-2010-054 and by the Grant Agency of Academy of Sciences of the Czech Republic., project. No. 1ET101470416

data) faces the issues of possibly incompatible lemma sets in various systems prompted us to try to evaluate the performance of different lemmatizers extrinsically - by measuring their effect on the results of another task, in this case the information retrieval. Furthermore, we also wanted to test our hypothesis that the quality of our automatically trained lemmatizer (measured through the means of IR performance) is fully comparable with the quality of the lemmatizer employing carefully prepared handcrafted dictionary, even thought the intrinsic performance measures suggest the superiority of the latter system. If such hypothesis is corroborated, it would hint that the researchers who would be developing lemmatizers for IR purposes in new languages do not have to implement a perfect handcrafted lemmatizer but could rely on the automatically trained one whose development is much faster.

2 Description of Lemmatizers

There are two main processes used for derivation of new words in a language: the inflectional and the derivative process. The words are derived from the same morphological class (for example the form *cleared* and *clears* of the verb *clear*) in the inflectional process while in the derivative process are derived from other morphological classes (*clearly*). The creation of a new word can be reached by applying a set of derivation rules in the both processes. The rules provide adding or stripping prefixes (prefix rule) and suffixes (suffix rule) to derive a new word form. From this point of view, the lemmatization can be regarded as the inverse operation to the inflectional and derivative processes.

We will compare two different approaches (manual versus automatic) to the lemmatizer construction and its influence on IR system in our experiments. For this purpose we use two different lemmatizers. The first one is based on the handcrafted dictionary of lemmas and set of affix (prefix and suffix) patterns. The second one is automatically trained lemmatizer.

2.1 Handcrafted Lemmatizer

The state-of-the-art Czech morphological analyzer which is available as part of The Prague Dependency Treebank[1] [1] was selected as a representative of handcrafted lemmatizers. The analyzer provides all possible lemmas for a given word and also a set of all conceivable morphological tags. The analyzer uses the handcrafted dictionary of lemmas (228,000 [2]) and manually created set of affix patterns (As is the author's best knowledge).

2.2 Automatically Trained Lemmatizer

Automatically created lemmatizer employed in our experiments is a slightly modified version of the lemmatizer introduced in [3]. This lemmatizer uses a

[1] We used the tool from version 2.0 of the treebank concretely.

dictionary of lemmas and a set of affix rules, both automatically inferred from training data. The training data consist of (full word form - lemma) pairs. The inference of lemmatization rules is based on searching for the longest common substring of the full form and the lemma. The lemmatization rules are in the form of if-then rules (for example, a simple lemmatization rule is: if a word ends by E, then strip E and add ION, i.e. in the symbolic form: $E > -E, ION$).

The main modification of the lemmatizer involves adding new patterns for lemmatization of out-of-vocabulary (OOV) words, that is, the word forms that were not seen in the training dictionary. There are actually two types of OOVs — the ones whose lemma is missing in the training data as well and the ones whose lemma occurs in the training set but not in pair with the word form in question. The new patterns for OOV words arise by concatenation of particular prefix and suffix pattern for each pair (full word form - lemma) in the training data. In the previous version of the lemmatizer, the prefix and suffix patterns were used separately. So now the lemmatization of unknown words involves the creation of prefix and suffix patterns (the chain of applicable prefix or suffix rules) and its concatenation. Then the concatenated pattern is firstly searched in the pattern library and if it does not exist then particular prefix and suffix patterns are searched in the library. If the particular pattern is found in the library then a most probable rule associated with this patterns is used to process the given word.

2.3 Training Data and Comparison of Lemmatizers

Data from two different sources were used for training of the automatically constructed lemmatizer. The first source was the previously mentioned Prague Dependency Treebank 2.0 (PDT). The second one was the Czech dictionary of lemmas and derivation rule file from the spell-checking program Ispell [4]. PDT contains full word forms and the corresponding lemmas, thus the training data were obtained by simply extracting these pairs. The second set of training data was prepared from Ispell files by using our own morphological generator (the Ispell files contain rules that allow to generate all full word forms for each lemma in the dictionary). In Tab. 1 are the quantitative informations about both acquired training data sets (PDT and Ispell) and in Tab. 2 are the informations about automatically created lemmatizers (Lem_PDT and Lem_Ispell).

Comparison of the accuracy of different lemmatizers is a difficult task due to the need for the manually lemmatized test data. In addition, the evaluation of results should be done manually as well because different lemmatizers generally do not share the same set of lemmas. Strictly speaking, the lemma is usually the infinitive for the verbs and the word in masculine, singular and nominative form for other inflected part-of-speech types, but generally each word form can be chosen as lemma for the group of words with the same stem. This selection heavily depends on the decision made by the dictionary author or the training data annotator. We have proposed an indirect comparison of the lemmatizers through set of IR experiments for these reasons which will be described in the next section.

Table 1. PDT and Ispell training data.

Training set	# pairs	# lemmas
PDT	200 431	66 401
Ispell	4 315 161	297 701

Table 2. Automatically created lemmatizers.

	Lem_PDT	Lem_Ispell
# lemmas	66 401	297 701
# rules	2 431	2 683
# P rules	213	55
# S rules	2 218	2 628
# patterns	28 867	34 999
# P+S patterns	26 436	32 3331
# P patterns	213	55
# S patterns	2 218	2 613

The direct comparison is of course possible (and often performed) when the lemmatizers do share the same set of lemmas. Since this is the case of the handcrafted lemmatizer (Lem_H) and the lemmatizer trained on PDT training data (Lem_PDT), we have compared them directly and the results are in Tab. 3. Recall (the number of the correctly lemmatized words to the number of all processed words ratio) (R), precision (the ratio of the number of the correctly lemmatized words to the number of all lemmas generated by the lemmatizer for all correctly lemmatized words) (P) and a harmonic F-measure ($(2 \cdot R \cdot P)/(R+P)$) ($F$) were evaluated on the test data part of the PDT corpus (the train and the development part were used for the lemmatizer training). The label Lem_H_G denotes the handcrafted lemmatizer with morphological guesser turned on (the guesser does not try to guess the correct lemma but only all possible morphological tags and, in addition, produces all presumably valid word forms for a given word). The labels Lem_PDT_oP and Lem_PDT_min denote the automatically trained lemmatizer using only OOV word patterns for lemmatization of all given words and the automatically trained lemmatizer using only prefix and suffix OOV patterns, respectively. No dictionary is used for lemma searching in the latter case and therefore this configuration can be seen as the minimal lemmatizer. In three last columns of the table are the results for the lemmatization of OOV words (words reported as unknown by Lem_H). There is only a small difference between both lemmatizers (Lem_H_G and Lem_PDT) recall (0.4 %) while the gap between precisions is much more significant (6.06 %). We will investigate the influence of these differences on the IR system in the next section.

3 IR Experiments

As mentioned before, our goal was to compare two approaches to the lemmatization on a real problem. Lemmatization was shown to improve the effectiveness

Table 3. Comparison of the lemmatizers.

	Test data			OOV words		
	R[%]	P[%]	F	R[%]	P[%]	F
Lem_H	99.38	82.45	0.90	73.41	100.00	0.85
Lem_H_G	99.50	79.71	0.89	93.88	12.90	0.23
Lem_PDT	99.10	73.65	0.85	75.35	96.19	0.85
Lem_PDT_oP	81.77	98.33	0.89	73.26	99.09	0.84
Lem_PDT_min	75.79	98.59	0.86	72.67	99.69	0.84

of information retrieval in highly inflected languages (as is the Czech language) in earlier experiments [5], [6].

3.1 Experimental Data

Our IR experiments were performed on the IR collection that was used in the Czech task of the Cross-Language Speech Retrieval track organized within the CLEF 2007 evaluation campaign[7]. This collection contains automatically transcribed spontaneous interviews (segmented by sliding a fixed-size window over the transcribed text into "documents") and two sets of TREC-like topics - 29 training and 42 evaluation topics. Each topic consists of 3 fields - `<title>` (T), `<desc>` (D) and `<narr>` (N).

Both sets of topics were used for the experiments and two types of queries were created for each set of topics - first one from the terms from the T and D fields and the second one from all terms from the fields T, D and N. Stop words were omitted from all sets of query terms. The aforementioned mGAP measure that was used in the CLEF 2007 Czech task was used as an evaluation measure.

The correct lemma for our experiments is chosen based on the disambiguation of the output of the morphological analyzer by a tagger for the Lem_H_G lemmatizer whereas for the automatically trained lemmatizer the first supplied lemma is chosen.

3.2 IR System

Language modeling approach [8] was used as the information retrieval method for the lemmatizer evaluation, concretely the query likelihood method with an linear interpolation unigram language model of the document with an unigram language model of the whole collection. The idea of this method is to create a language model M_d from each document d and then for each query q to find the model which most likely generated the query, that means to rank the documents according to the probability $P(d|q)$. We use the Bayes rule: $P(d|q) = P(q|d)P(d)/P(q)$, where $P(q)$ is the same for all documents and the prior document probability $P(d)$ is uniform across all documents, so we can ignore both. We have left the probability of the query been generated by a document model $P(q|M_d)$, which can be estimated using the maximum likelihood

Table 4. Comparison of mGAP score for lemmatized and non-lemmatized queries

test data	words	Lem_H_G	Lem_PDT	Lem_Ispell
train TD	0.0163	0.0270	0.0322	0.0280
train TDN	0.0164	0.0343	0.0364	0.0362
eval TD	0.0114	0.0220	0.0250	0.0200
eval TDN	0.0126	0.0274	0.0307	0.0243

estimate (MLE): $\hat{P}(q|M_d) = \prod_{t \in q} \frac{tf_{t,d}}{L_d}$, where $tf_{t,d}$ is the frequency of the term t in d and L_d is the total number of tokens in d. To deal with the sparse data for the generation of the M_d we use the mixture model between the document-specific multinomial distribution and the multinomial distribution of the whole collection M_c with interpolation parameter λ . So the final equation for ranking the documents according to the query is: $P(d|q) \propto \prod_{t \in q}(\lambda P(t|M_d) + (1-\lambda)P(t|M_c))$

3.3 Experiments Results

In the following text we compare the retrieval results of the two approaches to the lemmatization described above. For the case of automatically created lemmatizer we have two sets of results - each for different lemmatizer training data. Table 4 shows the mGAP score for the two sets of test data (training, evaluation) and the two sets of terms (TD, TDN) as described in Sect. 3.1. Interpolation parameter λ was set to 0.5. The retrieval results for all three lemmatizers are significantly better than the result for non-lemmatized data (words) for all sets of queries and terms.

As can be seen from table 4, the retrieval results when compared with Lem_H_G lemmatizer are better for the Lem_PDT lemmatizer for both sets of queries and terms. For the Lem_Ispell (again compared with Lem_H_G) the results are better for the training set of queries and worse for the evaluation set. Because the retrieval performance of this IR system can differ for various levels of interpolation, we have run tests for few different settings of λ. The results are shown in tables 5 and 6, pretty similar course for all levels of interpolation can be seen there.

3.4 Results evaluation

For the confirmation of our hypotheses, we ran several statistical significance tests. First, we claim that the retrieval results for the lemmatized data are better than the results for non-lemmatized data. The difference has shown to be statistically significant (with the significance level $\alpha = 0.01$) for all three tested lemmatizers when tested across all the query and terms sets and different settings of the retrieval method. The difference has also shown to be statistically significant when tested across the queries in one set for one setting of the IR method.

Table 5. Comparison of mGAP score for lemmatized queries of training set

term set	TD					TDN				
lemma / λ	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
Lem_H_G	0.0306	0.0290	0.0270	0.0261	0.0251	0.0392	0.0376	0.0343	0.0317	0.0295
Lem_PDT	0.0352	0.0343	0.0322	0.0298	0.0278	0.0396	0.0388	0.0364	0.0343	0.0307
Lem_Ispell	0.0328	0.0303	0.0280	0.0268	0.0255	0.0415	0.0397	0.0362	0.0329	0.0306
Lem_PDT_min	0.0326	0.0321	0.0305	0.0277	0.0264	0.0364	0.0345	0.0325	0.0305	0.0269
Lem_Ispell_min	0.0286	0.0274	0.0255	0.0231	0.0221	0.0394	0.0374	0.0347	0.0321	0.0296

Table 6. Comparison of mGAP score for lemmatized queries of evaluation set

term set	TD					TDN				
lemma / λ	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
Lem_H_G	0.0200	0.0212	0.0220	0.0222	0.0215	0.0255	0.0257	0.0274	0.0271	0.0260
Lem_PDT	0.0236	0.0243	0.0250	0.0252	0.0250	0.0281	0.0310	0.0307	0.0287	0.0271
Lem_Ispell	0.0193	0.0200	0.0200	0.0194	0.0198	0.0227	0.0234	0.0243	0.0243	0.0235
Lem_PDT_min	0.0186	0.0193	0.0197	0.0198	0.0195	0.0217	0.0215	0.0219	0.0215	0.0209
Lem_Ispell_min	0.0192	0.0199	0.0205	0.0204	0.0197	0.0178	0.0185	0.0178	0.0181	0.0168

Then we tested automatically created lemmatizers against the manually created one. When tested across all the query and terms sets and different settings of IR method, the difference between Lem_H_G and Lem_PDT has shown to be statistically significant (with the significance level $\alpha = 0.01$) and the difference between Lem_H_G and Lem_Ispell has not shown to be statistically significant. When tested across queries in one set the difference for both automatically created lemmatizers has not shown to be statistically significant. We believe that is due to the large variance of the GAP score among the queries in the set and small number of queries. The Wilcoxon Matched-Pairs Signed-Ranks Test [9] was used for all tests. The last two rows in tables 5 and 6 show retrieval results for lemmatizers with minimal configuration (Lem_PDT_min, Lem_Ispell_min). The difference in the recall of the lemmatizers seems to affect the retrieval precision, but the result is still superior in comparison with using non-lemmatized data and is especially suitable for the memory efficient IR systems.

4 Conclusions and future work

The results achieved in experiments shown in Sec. 3.3 suggest that, when using the lemmatizer for the IR system purposes, there is no substantial difference in performance between manually and automatically created lemmatizer. Actually, the automatically created lemmatizer (Lem_PDT) even improved the retrieval performance within our experimental setting (as the gain in the mGAP score has been shown to be statistically significant for the IR paradigm and the test

collection we have used - see Sect. 3.4). This result is especially promising in the prospect of development of IR systems for other languages since thanks to the existence of the Ispell resources for many languages, an acceptable lemmatizer can be easily built without any need of a manually created corpora or a handcrafted morphological analyzer (lemmatizer).

Just based on the presented experiments, it can not be said for sure what caused the observed performance gain. The first analysis of the results hints that the improvement could stem from the different approach to the lemmatization of some terms crucial for retrieving the relevant documents rather than from better overall precision and/or recall of the lemmatizer. More thorough examination of these causes and also a large-scale testing of these phenomenons using other information retrieval methods is a suitable matter for further work.

References

1. Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.: Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia, USA (2006)
2. Hajič, J., Hladká, B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of COLING-ACL Conference, Montreal, Canada (1998) 483–490
3. Kanis, J., Müller, L.: Automatic lemmatizer construction with focus on OOV words lemmatization. In: Proceedings of TSD 2005. Lecture Notes in Artificial Intelligence, Carlsbad, Czech Republic (2005) 132–139
4. Ispell dictionaries and rules files: fmg-www.cs.ucla.edu/geoff/ispell-dictionaries.html
5. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Lecture Notes in Computer Science, Alicante, Spain (2007) 759–765
6. Ircing, P., Psutka, J., Vavruška, J.: What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods UWB at CLEF 2007 CL-SR Track. In: Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, Budapest, Hungary (2008) 712–718
7. Ircing, P., Pecina, P., Oard, D.W., Wang, J., White, R.W., Hoidekr, J.: Information Retrieval Test Collection for Searching Spontaneous Czech Speech. In: Proceedings of TSD 2007. Lecture Notes in Artificial Intelligence, Plzeň, Czech Republic (2007) 439–446
8. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1998) 275–281
9. The Wilcoxon matched-pairs signed-ranks test: www.fon.hum.uva.nl/service/statistics/signed_ rank_ test.html