Lemmatisation as a Tagging Task

Andrea Gesmundo
Department of Computer Science
University of Geneva
andrea.gesmundo@unige.ch

Tanja Samardžić
Department of Linguistics
University of Geneva
tanja.samardzic@unige.ch

Abstract

We present a novel approach to the task of word lemmatisation. We formalise lemmatisation as a category tagging task, by describing how a word-to-lemma transformation rule can be encoded in a single label and how a set of such labels can be inferred for a specific language. In this way, a lemmatisation system can be trained and tested using any supervised tagging model. In contrast to previous approaches, the proposed technique allows us to easily integrate relevant contextual information. We test our approach on eight languages reaching a new state-of-the-art level for the lemmatisation task.

1 Introduction

Lemmatisation and part-of-speech (POS) tagging are necessary steps in automatic processing of language corpora. This annotation is a prerequisite for developing systems for more sophisticated automatic processing such as information retrieval, as well as for using language corpora in linguistic research and in the humanities. Lemmatisation is especially important for processing morphologically rich languages, where the number of different word forms is too large to be included in the part-of-speech tag set. The work on morphologically rich languages suggests that using comprehensive morphological dictionaries is necessary for achieving good results (Hajič, 2000; Erjavec and Džeroski, 2004). However, such dictionaries are constructed manually and they cannot be expected to be developed quickly for many languages.

In this paper, we present a new general approach to the task of lemmatisation which can be used to overcome the shortage of comprehensive dictionaries for languages for which they have not been developed. Our approach is based on redefining the task of lemmatisation as a category tagging task. Formulating lemmatisation as a tagging task allows the use of advanced tagging techniques, and the efficient integration of contextual information. We show that this approach gives the highest accuracy known on eight European languages having different morphological complexity, including agglutinative (Hungarian, Estonian) and fusional (Slavic) languages.

2 Lemmatisation as a Tagging Task

Lemmatisation is the task of grouping together word forms that belong to the same inflectional morphological paradigm and assigning to each paradigm its corresponding canonical form called lemma. For example, English word forms *go*, *goes*, *going*, *went*, *gone* constitute a single morphological paradigm which is assigned the lemma *go*. Automatic lemmatisation requires defining a model that can determine the lemma for a given word form. Approaching it directly as a tagging task by considering the lemma itself as the tag to be assigned is clearly unfeasible: 1) the size of the tag set would be proportional to the vocabulary size, and 2) such a model would overfit the training corpus missing important morphological generalisations required to predict the lemma of unseen words (e.g. the fact that the transformation from *going* to *go* is governed by a general rule that applies to most English verbs).

Our method assigns to each word a label encod-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 368–372,
Jeju, Republic of Korea, 8-14 July 2012.

ing the transformation required to obtain the lemma string from the given word string. The generic transformation from a word to a lemma is done in four steps: 1) remove a suffix of length N_s; 2) add a new lemma suffix, L_s; 3) remove a prefix of length N_p; 4) add a new lemma prefix, L_p. The tuple $\tau \equiv \langle N_s, L_s, N_p, L_p \rangle$ defines the word-to-lemma transformation. Each tuple is represented with a label that lists the 4 parameters. For example, the transformation of the word *going* into its lemma is encoded by the label $\langle 3, \emptyset, 0, \emptyset \rangle$. This label can be observed on a specific lemma-word pair in the training set but it generalizes well to the unseen words that are formed regularly by adding the suffix *-ing*. The same label applies to any other transformation which requires only removing the last 3 characters of the word string.

Suffix transformations are more frequent than prefix transformations (Jongejan and Dalianis, 2009). In some languages, such as English, it is sufficient to define only suffix transformations. In this case, all the labels will have N_p set to 0 and L_p set to $\emptyset$. However, languages richer in morphology often require encoding prefix transformations too. For example, in assigning the lemma to the negated verb forms in Czech the negation prefix needs to be removed. In this case, the label $\langle 1, t, 2, \emptyset \rangle$ maps the word *nevěděl* to the lemma *vědět*. The same label generalises to other (word, lemma) pairs: (*nedokázal*, *dokázat*), (*neexistoval*, *existovat*), (*nepamatoval*, *pamatovat*).[1]

The set of labels for a specific language is induced from a training set of pairs (word, lemma). For each pair, we first find the Longest Common Substring (LCS) (Gusfield, 1997). Then we set the value of N_p to the number of characters in the word that precede the start of LCS and N_s to the number of characters in the word that follow the end of LCS. The value of L_p is the substring preceding LCS in the lemma and the value of L_s is the substring following LCS in the lemma. In the case of the example pair (*nevěděl*, *vědět*), the LCS is *vědě*, 2 characters precede the LCS in the word and 1 follows it. There are no characters preceding the start of the LCS in the lemma and '*t*' follows it. The generated label is added to the set of labels.

[1]The transformation rules described in this section are well adapted for a wide range of languages which encode morphological information by means of affixes. Other encodings can be designed to handle other morphological types (such as Semitic languages).

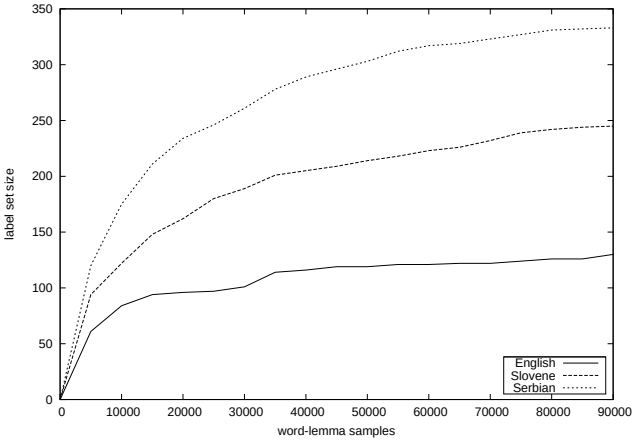

Figure 1: Growth of the label set with the number of training instances.

3 Label set induction

We apply the presented technique to induce the label set from annotated running text. This approach results in a set of labels whose size convergences quickly with the increase of training pairs.

Figure 1 shows the growth of the label set size with the number of tokens seen in the training set for three representative languages. This behavior is expected on the basis of the known interaction between the frequency and the regularity of word forms that is shared by all languages: infrequent words tend to be formed according to a regular pattern, while irregular word forms tend to occur in frequent words. The described procedure leverages this fact to induce a label set that covers most of the word occurrences in a text: a specialized label is learnt for frequent irregular words, while a generic label is learnt to handle words that follow a regular pattern.

We observe that the non-complete convergence of the label set size is, to a large extent, due to the presence of noise in the corpus (annotation errors, typos or inconsistency). We test the robustness of our method by deciding not to filter out the noise generated labels in the experimental evaluation. We also observe that encoding the prefix transformation in the label is fundamental for handling the size of the label sets in the languages that frequently use lemma prefixes. For example, the label set generated for

Czech doubles in size if only the suffix transformation is encoded in the label. Finally, we observe that the size of the set of induced labels depends on the morphological complexity of languages, as shown in Figure 1. The English set is smaller than the Slovene and Serbian sets.

4 Experimental Evaluation

The advantage of structuring the lemmatisation task as a tagging task is that it allows us to apply successful tagging techniques and use the context information in assigning transformation labels to the words in a text. For the experimental evaluations we use the Bidirectional Tagger with Guided Learning presented in Shen et al. (2007). We chose this model since it has been shown to be easily adaptable for solving a wide set of tagging and chunking tasks obtaining state-of-the-art performances with short execution time (Gesmundo, 2011). Furthermore, this model has consistently shown good generalisation behaviour reaching significantly higher accuracy in tagging unknown words than other systems.

We train and test the tagger on manually annotated G. Orwell's "1984" and its translations to seven European languages (see Table 2, column 1), included in the Multext-East corpora (Erjavec, 2010). The words in the corpus are annotated with both lemmas and detailed morphosyntactic descriptions including the POS labels. The corpus contains 6737 sentences (approximatively $110k$ tokens) for each language. We use 90% of the sentences for training and 10% for testing.

We compare lemmatisation performance in different settings. Each setting is defined by the set of features that are used for training and prediction. Table 1 reports the four feature sets used. Table 2 reports the accuracy scores achieved in each setting. We establish the Base Line (BL) setting and performance in the first experiment. This setting involves only features of the current word, $[w_0]$, such as the word form, suffixes and prefixes and features that flag the presence of special characters (digits, hyphen, caps). The BL accuracy is reported in the second column of Table 2).

In the second experiment, the BL feature set is expanded with features of the surrounding words ($[w_{-1}]$, $[w_1]$) and surrounding predicted lemmas ($[lem_{-1}]$, $[lem_1]$). The accuracy scores obtained in the second experiment are reported in the third column of Table 2. The consistent improvements over the BL scores for all the languages, varying from the lowest relative error reduction (RER) for Czech (5.8%) to the highest for Romanian (31.6%), confirm the significance of the context information. In the third experiment, we use a feature set in which the BL set is expanded with the predicted POS tag of the current word, $[pos_0]$.[2] The accuracy measured in the third experiment (Table 2, column 4) shows consistent improvement over the BL (the best RER is 34.2% for Romanian). Furthermore, we observe that the accuracy scores in the third experiment are close to those in the second experiment. This allows us to state that it is possible to design high quality lemmatisation systems which are independent of the POS tagging. Instead of using the POS information, which is currently standard practice for lemmatisation, the task can be performed in a context-wise setting using only the information about surrounding words and lemmas.

In the fourth experiment we use a feature set consisting of contextual features of words, predicted lemmas and predicted POS tags. This setting com-

Base Line (BL)	$[w_0]$, *flagChars*(w_0), *prefixes*(w_0), *suffixes*(w_0)
+ context	BL + $[w_1]$, $[w_{-1}]$, $[lem_1]$, $[lem_{-1}]$
+ POS	BL + $[pos_0]$
+cont.&POS	BL + $[w_1]$, $[w_{-1}]$, $[lem_1]$, $[lem_{-1}]$, $[pos_0]$, $[pos_{-1}]$, $[pos_1]$

Table 1: Feature sets.

	Base	+	+	+cont.&POS	
Language	Line	cont.	POS	Acc.	UWA
Czech	96.6	96.8	96.8	97.7	86.3
English	98.8	99.1	99.2	99.6	94.7
Estonian	95.8	96.2	96.5	97.4	78.5
Hungarian	96.5	96.9	97.0	97.5	85.8
Polish	95.3	95.6	96.0	96.8	85.8
Romanian	96.2	97.4	97.5	98.3	86.9
Serbian	95.0	95.3	96.2	97.2	84.9
Slovene	96.1	96.6	97.0	98.1	87.7

Table 2: Accuracy of the lemmatizer in the four settings.

[2]The POS tags that we use are extracted from the morphosyntactic descriptions provided in the corpus and learned using the same system that we use for lemmatisation.

bines the use of the context with the use of the predicted POS tags. The scores obtained in the fourth experiment are considerably higher than those in the previous experiments (Table 2, column 5). The RER computed against the BL varies between 28.1% for Hungarian and 66.7% for English. For this setting, we also report accuracies on unseen words only (UWA, column 6 in Table 2) to show the generalisation capacities of the lemmatizer. The UWA scores 85% or higher for all the languages except Estonian (78.5%).

The results of the fourth experiment show that interesting improvements in the performance are obtained by combining the POS and context information. This option has not been explored before. Current systems typically use only the information on the POS of the target word together with lemmatisation rules acquired separately from a dictionary, which roughly corresponds to the setting of our third experiment. The improvement in the fourth experiment compared to the third experiment (RER varying between 12.5% for Czech and 50% for English) shows the advantage of our context-sensitive approach over the currently used techniques.

All the scores reported in Table 2 represent performance with raw text as input. It is important to stress that the results are achieved using a general tagging system trained only a small manually annotated corpus, with no language specific external sources of data such as independent morphological dictionaries, which have been considered necessary for efficient processing of morphologically rich languages.

5 Related Work

Juršič et al. (2010) propose a general multilingual lemmatisation tool, LemGen, which is tested on the same corpora that we used in our evaluation. LemGen learns word transformations in the form of *ripple-down rules*. Disambiguition between multiple possible lemmas for a word form is based on the gold-standard morphosyntactic label of the word. Our system outperforms LemGen on all the languages. We measure a Relative Error Reduction varying between 81% for Serbian and 86% for English. It is worth noting that we do not use manually constructed dictionaries for training, while Juršič et al. (2010) use additional dictionaries for languages for which they are available.

Chrupała (2006) proposes a system which, like our system, learns the lemmatisation rules from a corpus, without external dictionaries. The mappings between word forms and lemmas are encoded by means of the *shortest edit script*. The sets of edit instructions are considered as class labels. They are learnt using a SVM classifier and the word context features. The most important limitation of this approach is that it cannot deal with both suffixes and prefixes at the same time, which is crucial for efficient processing of morphologically rich languages. Our approach enables encoding transformations on both sides of words. Furthermore, we propose a more straightforward and a more compact way of encoding the lemmatisation rules.

The majority of other methods are concentrated on lemmatising out-of-lexicon words. Toutanova and Cherry (2009) propose a joint model for assigning the set of possible lemmas and POS tags to out-of-lexicon words which is language independent. The lemmatizer component is a discriminative character transducer that uses a set of within-word features to learn the transformations from input data consisting of a lexicon with full morphological paradigms and unlabelled texts. They show that the joint model outperforms the pipeline model where the POS tag is used as input to the lemmatisation component.

6 Conclusion

We have shown that redefining the task of lemmatisation as a category tagging task and using an efficient tagger to perform it results in a performance that is at the state-of-the-art level. The adaptive general classification model used in our approach makes use of different sources of information that can be found in a small annotated corpus, with no need for comprehensive, manually constructed morphological dictionaries. For this reason, it can be expected to be easily portable across languages enabling good quality processing of languages with complex morphology and scarce resources.

7 Acknowledgements

The work described in this paper was partially funded by the Swiss National Science Foundation grants CRSI22 127510 (COMTIS) and 122643.

References

Grzegorz Chrupała. 2006. Simple data-driven context-sensitive lemmatization. In *Proceedings of the Sociedad Española para el Procesamiento del Lenguaje Natural*, volume 37, page 121131, Zaragoza, Spain.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: lemmatizing unknown Slovene words. *Applied Artificial Intelligence*, 18:17–41.

Tomaž Erjavec. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2544–2547, Valletta, Malta. European Language Resources Association (ELRA).

Andrea Gesmundo. 2011. Bidirectional sequence classification for tagging tasks with guided learning. In *Proceedings of TALN 2011*, Montpellier, France.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press.

Jan Hajič. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 94–101, Seattle, Washington. Association for Computational Linguistics.

Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153, Suntec, Singapore, August. Association for Computational Linguistics.

Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. LemmaGen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767, Prague, Czech Republic. Association for Computational Linguistics.

Kristina Toutanova and Colin Cherry. 2009. A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, page 486494, Suntec, Singapore. Association for Computational Linguistics.