TLT-CRF: A Lexicon-supported Morphological Tagger for Latin Based on Conditional Random Fields

Tim vor der Brück, Alexander Mehler
Distributed Secure Software Systems, Text Technology Lab
Lucerne University of Applied Sciences and Arts, Goethe-Universität Frankfurt am Main
tim.vorderbrueck@hslu.ch, amehler@em.uni-frankfurt.de

Abstract

We present a morphological tagger for Latin, called *TTLab Latin Tagger based on Conditional Random Fields* (TLT-CRF) which uses a large Latin lexicon. Beyond Part of Speech (PoS), TLT-CRF tags eight inflectional categories of verbs, adjectives or nouns. It utilizes a statistical model based on CRFs together with a rule interpreter that addresses scenarios of sparse training data. We present results of evaluating TLT-CRF to answer the question what can be learnt following the paradigm of 1st order CRFs in conjunction with a large lexical resource and a rule interpreter. Furthermore, we investigate the contigency of representational features and targeted parts of speech to learn about selective features.

Keywords: morphological tagging, rules, CRFs, Latin lexicon

1. Introduction

A tagger that determines morphological, inflectional information for each word beyond its PoS is called a *morphological tagger*. We describe a morphological tagger (the tagset being employed is a subset of the *Stuttgart-Tübingen TagSet* (STTS)) called *TTLab Latin Tagger based on Conditional Random Fields* (TLT-CRF). It determines the PoS as well as inflectional categories such as *number*, *case*, *gender*, *comparison degree*, *mood*, *tense* and *voice*. TLT-CRF is based on a CRF model together with a large Latin lexicon. It explores several lexical features (e.g., for determining agreement) and, thus, is distinguished from related approaches. To the best of our knowledge, it is the only hybrid morphological tagger that integrates a rule interpreter and a statistical model in the area of Latin texts. Rules are useful in cases of data sparseness which do not allow for defining sufficient training data. The source code of the tagger can be obtained from `http://prepro.hucompute.org`.

2. Related Work

There is a multitude of statistical approaches for PoS tagging. Almost all of them rely on a statistical model that is trained on annotated examples and then applied to previously unseen texts. These models include decision trees (Schmid, 1994), Hidden Markov Models (Brants, 2000), Maximum Entropy Models (Toutanova et al., 2003), neural networks and structured SVMs. An exception is described by (Bellegarda, 2010) who employs *learning per analogy* together with latent semantic analysis (Landauer et al., 1998) (called *latent analogy*). While there exists a lot of freely available PoS taggers, so-called morphological taggers are still rare. An exception is (Müller and Schütze, 2015) who provide a morphological tagger for several languages that is based on 3rd order CRFs (Lafferty et al., 2001). A hybrid morphological tagger for Czech is introduced by (Spoustová, 2008). The employed statistical model is a combination of HMMs, of a maximum entropy model and of perceptrons, where the first two models are now seen to be outdated due to problems of including arbitrary features (HMM) and due to the label bias problem (maximum entropy models) (Sutton and McCallum, 2007). Another approach to utilizing state-of-the art taggers is proposed by (Eger et al., 2016) mostly based on MarMoT (Müller and Schütze, 2015) and, thus, on higher-order CRFs. In contrast to this, we focus on 1st order CRFs by asking about the payout of such a simpler model. To know whether 1st order CRFs allow for competitive alternatives is important for scenarios in which time is a critical variable since more complex models require longer training and processing times.

3. Lexical Rules

TLT-CRF can handle manually generated tagging rules as a special class of features. Such rules are used to specify the PoS or lemma of an input token subject to a set of conditions to be met by the token and its context in the input text. Conditions are defined by means of regular expressions over inflectional categories and wordforms

in the lexical context of the input token. Rules can either include sufficient or necessary conditions. Rules based on sufficient conditions determine the PoS/lemma of a token which meets their conditions. In contrast to this, any PoS is ruled out that is associated with a failing rule based on necessary conditions. Since there is practically no rule without exceptions, the tagger allows for considering rules in a soft mode. In this mode, TLT-CRF creates features using the names of the rules that apply to the focal token. The decision which PoS to chose is then left to the CRF model. An example of a tagging rule is shown below: it says that the PoS of *ut* is *conjunction*, if *ut* is (not necessarily directly) followed by a finite verb of mood *subjunctive*.

```
<rule name="ut">
 <premise type="sequence"
  forwards="true">
  <rule_element type="atom">
   <word_form>ut</word_form>
  </rule_element>
  <rule_element type="*">
   <rule_element type="atom">
    <pos negated="true">V</pos>
   </rule_element>
  </rule_element>
  <rule_element type="?">
   <rule_element type="atom">
    <pos>V</pos>
     <mood>INFINITIVE</mood>
   </rule_element>
  </rule_element>
  <rule_element type="atom">
   <pos>V</pos>
   <mood>SUBJUNCTIVE</mood>
  </rule_element>
 </premise>
 <conclusion pos="CON" />
</rule>
```

4. The Frankfurt Latin Lexicon

As a large Latin lexical resource of our tagger, we utilize the *Frankfurt Latin Lexicon* (FLL) (Mehler et al., 2015), which can be freely browsed via the website `http://collex.hucompute.org`. The FLL is based on an automatic morphological expansion of a range of lemmata extracted from web-based resources (e.g., LemLat (Passerotti, 2004), Perseus Digital Library (Smith et al., 2000), Whitaker's word list[1], the Latin Wiktionary[2], Latin training data of the Tree Tagger (Schmid, 1994)) and textual resources (e.g., the Patrologia Latina). In the meantime, more than 7% of all lemmas of the FLL have been manually created or checked by experts of Latin via the website of *Computational Historical Semantics* (Jussen et al., 2007): `http://www.comphistsem.org/home.html`. Since the FLL creates and stores all inflected forms of its lemmas, it now contains more than 12 million syntactic words (i.e., wordforms plus inflectional features like *person*, *number*, *case* etc.). This provides a huge resource for morphological tagging as addressed by TLT-CRF.

5. Corpus

Our evaluation and training corpus is based on the capitularies, the amalarius corpus and three further texts from the MGH[3] corpus (Visio Baronti, Vita Adelphii, Vita Amandi). The entire corpus is tokenized and split into sentences. Each token of the corpus is manually assigned a unique id that references the corresponding syntactic word within the FLL. In this way, full morphological information is available for all tokens of this corpus. By sampling (randomly rearranging) the sentences of this corpus, we make available this gold standard data.[4]

6. TLT-CRF

TLT-CRF is mainly based on a statistical model in the framework of *Conditional Random Fields* (CRF) (Lafferty et al., 2001). It uses the CRFsuite (Okazaki, 2007) which uses the limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) training method to implement a 1st order CRF. A selection of the features employed by TLT-CRF is as follows:

1. **Capitalization:** named entities and certain adjectives are usually capitalized in Latin texts. Further, in order to emphasize important words, authors may capitalize them (e.g., *Deus – God*). To account for such phenomena, we use a Boolean feature that is set to *true* iff the

[1] `http://archives.nd.edu/whitaker/dictpage.htm`

[2] `http://la.wiktionary.org/wiki/Pagina_prima`

[3] The acronym MGH stands for *Monumenta Germaniae Historica*; see `http://www.mgh.de`.

[4] Please contact the 2nd author of this paper or consult the `www.hucompute.org` website.

token begins in uppercase. This feature is not specified for sentence initial words.

2. **Suffix n-grams, $n = 1, \ldots, 3$:** suffixes often inform about grammatical features of the corresponding word and, thus, about its PoS. Thus, we extract suffixes as features using a CRF-based stemming method.
3. **Letter-based n-gram models:** we additionally apply a letter-based word-intrinsic n-gram model to predict PoS ($n = 1, \ldots, 4$).
4. **Lemma-based n-grams:** we look up all words in a left- and right-sided window around the focal token ($n = 1, \ldots, 3$). If the tokens in this window can be unambiguously mapped onto lemmas of the FLL, then the sequence of these lemmas is used as an additional feature.
5. **Numbers:** The feature `number` can take the values `CARDINAL_ARABIC` for numbers according to the Hindu-Arabic numeral system (1, 2, 3, ...), `CARDINAL_ROMAN` for Roman numerals (I, II, III, ...), `CARDINAL` for written-out numbers and `ORD` for ordinal numbers. Roman numerals are recognized by means of a regular expression.
6. **Word n-grams:** the PoS of a token usually depends on (the PoS of) its lexical context. This is a standard feature basically used by all PoS taggers.
7. **Skip-grams:** since Latin has a relatively free word order, we explore up to skip-trigrams of words in the context of the token to be tagged as additional features. For instance, nouns can follow or precede attributive adjectives.
8. **Agreement:** n-grams of gender, number and case are explored as features in order to account for grammatical agreement of nouns and adjectives.

Figure 1 shows the activity diagram of the entire tagging process of TLT-CRF and the different features being included. For lemmatization, we employ the lemmatizer *AliseTra* of (Eger, 2015). AliSeTra operates on single tokens disregarding their textual context. However, since this context can play an important role in choosing the correct lemma of a token, we combine the probability computed by AliSeTra with the one computed by a language model (SRILM – (Stolcke, 2002)) operating on lemma sequences. The total score of a candidate lemma is given by the arithmetic mean of both probabilities (normalized by the sum of the probabilities of all candidates). The morphological tagging of a token x is achieved by, firstly, tagging its PoS and the morphological categories independently using the same input features. In a second step, TLT-CRF identifies the syntactic word within the FLL that shares the same wordform and the majority of morphological categories assigned to x in the latter step. If such an entry exist, all its morphological features are finally assigned to x. Otherwise, the category values determined by the tagger in the first step are used. With this pipeline approach, we achieved a total morphological tagging accuracy of 88.08%. We improved this accuracy to 88.84% via partial joint tagging, which can also be accomplished with older CRF applications like CRFsuite that are not able to handle huge amount of tags. For that, besides of an isolated tagging of morphological labels, we jointly tag certain label combinations (currently only *case*, *gender*, and *number*). In case of ties in the above described lexicon lookup, we prefer the lexicon entries that are compliant with the result of the joint prediction.

6.1. Evaluation

We evaluate TLT-CRF with respect to lemmatization and morphological tagging using the Capitularies corpus (see above). We compare TLT-CRF with Lapos (Tsuruoka et al., 2011), TreeTagger (Schmid, 1994), StanfordTagger (Toutanova et al., 2003), OpenNLPTagger[5] and MarMoT (Müller and Schütze, 2015) by training and evaluating these taggers using the same data set. The results of our evaluation are shown in Table 2. According to these results, TLT-CRF achieves promising results in comparison to its state-of-the-art competitors. In our evaluation scenario, MarMoT is definitely the best performing tagger. However, TLT-CRF performs second best in the case of tagging PoS. MarMoT and TLT-CRF, which make use of the lexicon, perform best regarding PoS tagging. This stresses the importance of lexical resources. If we consider the inflectional categories in isolation, TLT-CRF performs worse than several of its competitors. However, if we consider joint learning of these inflectional categories, TLT-CRF performs second best again (see the line ALL in Table 2). This result shows that a 1st order CRF is a competitive alternative if a 3rd order CRF is out of reach (possibly due to restrictions of training time and processing time). Regarding the taggers of our evaluation, only MarMoT and TLT-CRF natively support the use of a lexicon (OpenNLP supports a lexicon but this use is not documented). See (Eger et al., 2015) who evaluate an approach to incorporating lexical knowledge into taggers without native lexicon

[5] `https://opennlp.apache.org/`

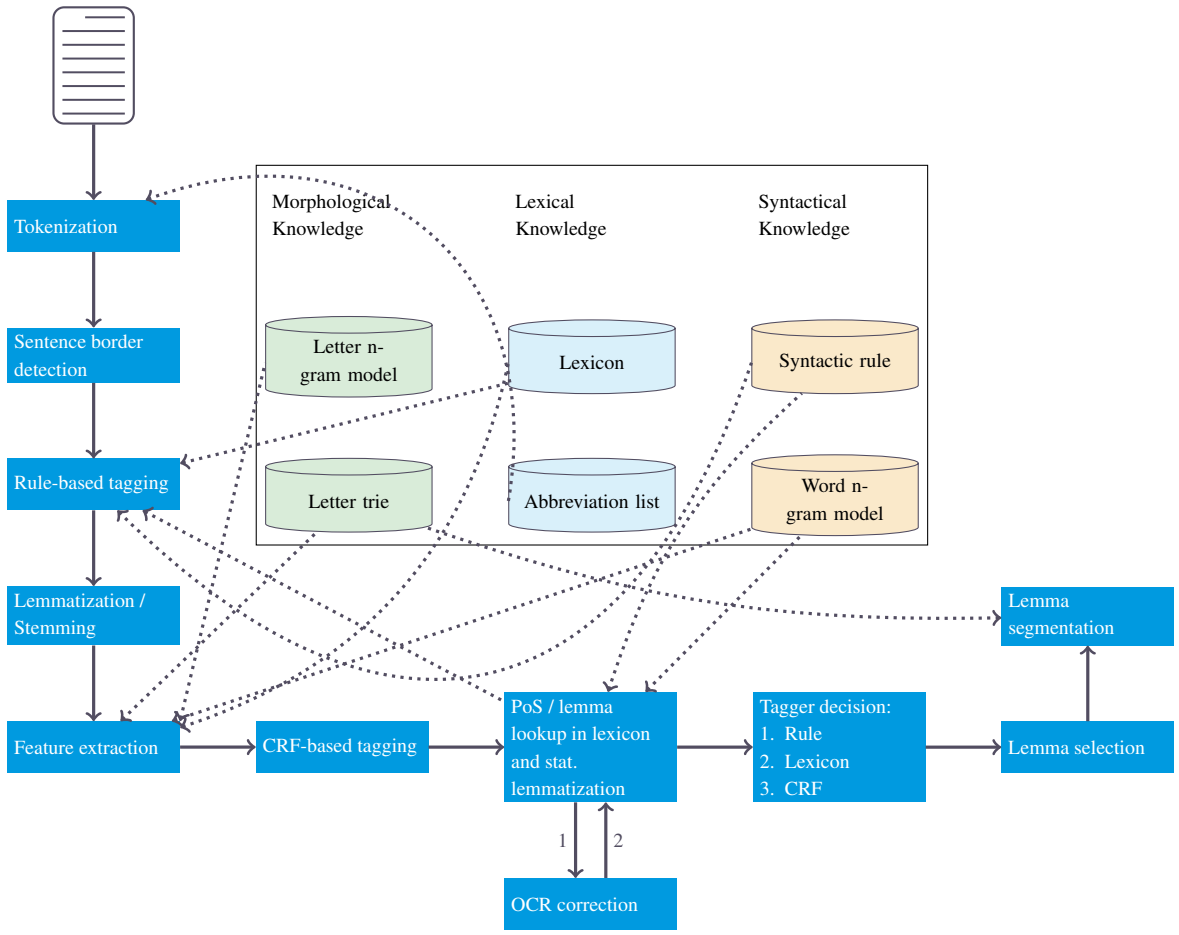

Figure 1: The activity diagram of TLT-CRF.

support.

In order to get more insights into tagging accuracy, we investigated the contingency of feature selection and PoS tagging, which may hint at the features' impact on predicting PoS. In particular, we calculated the χ^2 coefficient of representational features on the one hand and parts of speech on the other. To allow for comparison of features with different degrees of freedom, i.e., different number of feature values and associated PoS classes, we divide each χ^2 coefficient value by its critical value r for the significance level of 5%. Thus, a ratio above one indicates that feature and PoS are related (in terms of contingency). The most strongly correlated features are shown in Table 1. Numbers in brackets indicate the offsets to the focal token. The following feature abbreviations are used in Table 1:

- cap: focal token is capitalized or not
- intngram: PoS prediction of an intrinsic letter-based ngram model
- l: lemma as determined by a lexicon lookup (currently only unique lemmas are considered)
- ll: lemma as predicted by the lemmatizer
- lp: PoS associated to the lemma that is determined by the lemmatizer
- ls: suffix as determined by removing the lemma string, as determined by the lemmatizer, from the word form string
- number: token is a Roman, respectively an Hindu-Arabic numeral or not
- p: PoS from the lexicon
- r: last 7 characters of a word
- s: word skip-gram
- symbol: focal token is symbol (e.g., '.,!?:') or not
- w: word
- x: suffix as determined by the stemmer
- word length: discretized word length, can assume the values *short*, *medium*, *long*, *very long*

Feature	χ^2/r	χ^2
symbol	46063.795	432560.000
p[0]	29750.472	8428199.550
lp[0]	15474.089	3174283.217
cap[0]	14322.531	134495.0822
number[0]	9975.814	533330.301
cap[1]	7047.210	66176.505
word length	6731.752	256588.920
cap[2]	4522.818	42471.317
intngram	3023.860	503302.269
p[1]	1751.774	496271.014
p[-1]	1619.660	458843.664
number[-1]	1356.089	72499.680
lp[-1]	660.474	164915.485
p[2]	637.498	180600.883
number[1]	624.890	33408.0533
p[-2]	610.825	173044.363
cap[-2]	603.383	5666.042
lp[1]	597.684	149237.413
cap[-1]	432.626	4062.559
lp[2]	357.980	89385.080
lp[-2]	338.583	84541.795
x[0]	66.528	1532159.475
l[0]	28.887	4749664.203
pw[0]	27.448	6863222.365
ls[0]	25.507	124369.979
ll[0]	22.432	4838691.809
r[0]	12.617	6982236.483
x[1]	10.545	289541.849
w[0]	9.159	7070529.819
x[-1]	7.811	215592.093

Table 1: The 30 features that are most strongly correlated (in terms of contingency) to the PoS classes.

Note that *current word (w[0])* is not among the 20 most strongly correlated features, which is mainly caused by the fact that Latin words are highly ambiguous. For, instance, a lot of adverbs are homonymous with prepositions. The same holds for adjectives and nouns. The evaluation also shows that the features of strongest correlation are located in a context of words that are maximally two tokens away from the focal term. The rule feature is not among the 30 features most strongly correlated to the PoS, but its relative χ-squared value is till considerably above 1 (e.g., 1.38, absolute χ-square: 54.95). While the use of rules as hard constraints degrades PoS accuracy, there is a small gain when employing them in a soft mode (+0.01%). Finally, one can observe that an intrinsic n-gram model, which is barely used in state-of-the-art PoS taggers, can be quite beneficial.

7. Conclusion

We presented TLT-CRF as a hybrid morphological tagger for Latin which employs lexicon-based features as well as hand-crafted rules in the framework of 1st order CRFs. According to Table 2, TLT-CRF achieves promising results in comparison to several state-of-the-art competitors. TLT-CRF performs as the second-best tagger regarding PoS tagging and joint learning of inflectional categories. Because of its simple architecture that allows for including a wide range of additional morphological, lexical or syntactic features, it can be seen as a promising candidate for exploring feature-based models in the framework of 1st order CRFs. This finding is advantageous since CRFs of this sort are implemented very efficiently. Note that there are promising alternatives to the CRFsuite which we used for implementing TLT-CRF. Unlike MarMoT, for example, CRFsuite does not support higher order models and is quite inefficient in comparison to state of the art CRF systems. Thus, we plan to replace our current implementation of TLT-CRF by means of MarMoT while relying on the wide set of features studied in this paper.

Acknowledgment

We gratefully acknowledge financial support by the BMBF via the projects `www.comphistsem.org/project.html` and `www.cedifor.de/en`.

References

Bellegarda, J. R. (2010). Part-of-speech tagging by latent analogy. *IEEE Journal of Selected Topics in Signal Processing*, 4(6).

Brants, T. (2000). TnT: A statistical part-of-speech tagger. In *Proc. of the 6th Conf. on Applied Natural Language Processing*, ANLC '00, pages 224–231.

Eger, S., vor der Brück, T., and Mehler, A. (2015). Lexicon-assited tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods. In *Proc. of LaTeCH 2015*, pages 105–113, Bejing.

Eger, S., Gleim, R., and Mehler, A. (2016). Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proc. of LREC 2016*.

	TLT-CRF	MarMoT	Lapos	TnT	Mate	TreeTagger	Stanford	OpenNLP	
								Max.Entr	Perceptron
PoS	96.07	**96.32**	95.86	95.16	95.67	95.15	95.55	94.54	94.11
case	94.93	**95.51**	94.64	92.86	94.56	92.12	94.58	93.41	92.27
degree	97.22	**97.78**	97.55	97.09	97.40	97.03	97.30	96.65	96.19
gender	95.62	**96.72**	96.09	95.35	95.84	94.75	95.83	94.55	93.76
mood	97.19	**98.65**	98.28	97.73	98.13	97.73	98.12	97.14	97.16
number	97.48	**97.92**	97.19	96.91	97.04	96.61	97.23	96.03	95.44
person	99.24	**99.42**	99.25	98.87	99.27	98.89	99.18	98.53	98.40
tense	98.12	**98.88**	98.53	98.17	98.41	98.12	98.43	97.62	97.54
voice	98.79	**99.02**	98.79	98.52	98.74	98.55	98.67	98.30	98.23
ALL	88.84	**91.48**	88.19	85.73	87.66	84.53	85.28	83.38	80.96

Table 2: Tag accuracies in % for different systems and categories (cf. (Eger et al., 2015)). For non-morphological taggers, all inflectional categories and PoS are tagged in isolation. Note that we did not yet evaluate the Mate Morphological Tagger.

Eger, S. (2015). Designing and comparing g2p-type lemmatizers for a morphology-rich language. In *4th Int. Workshop on Systems and Frameworks for Computational Morphology*, Stuttgart.

Jussen, B., Mehler, A., and Ernst, A. (2007). A corpus management system for historical semantics. *Sprache und Datenverarbeitung. International Journal for Language Data Processing*, 31(1-2):81–89.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilitistic models for segmenting and labeling sequence data. In *Proc. of ICML 2001*, St. Petersburg/Russia.

Landauer, T., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25(2-–3):259—-284.

Mehler, A., vor der Brück, T., Gleim, R., and Geelhaar, T. (2015). Towards a network model of the coreness of texts: An experiment in classifying Latin texts using the TTLab Latin tagger. In Ch. Biemann et al., editors, *Text Mining: From Ontology Learning to Automated text Processing Applications*, pages 87–112. Springer, Berlin/New York.

Müller, T. and Schütze, H. (2015). Robust morphological tagging with word representations. In *Proc. of NAACL 2015*, pages 526–536, Denver. ACL.

Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Passerotti, M. (2004). Development and perspectives of the latin morphological analyser lemlat. *Linguistica Computazionale*, 20–21.

Schmid, H. (1994). Probabilistic parf-of-speech tagging using decision trees. In *Proc. of Int. Conf. on New Methods in Language Processing*, Manchester, UK.

Smith, D., Rydberg-Cox, J., and Crane, G. (2000). The perseus project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Spoustová, D. (2008). Combining statistical and rule-based approaches to morphological tagging of czech texts. *The Prague Bulletin of Mathematical Linguistics*, (89):23–40.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proc. of the International Confernce of Spoken Language Processing*, Denver, Colorado.

Sutton, C. and McCallum, A. (2007). An introduction to conditional random fields for relational learning. In Lise Getoor et al., editors, *Introduction to Statistical Relational Learning*. MIT Press.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL 2003*, pages 173–180, Stroudsburg, PA, USA.

Tsuruoka, Y., Miyao, Y., and Kazama, J. (2011). Learning with lookahead: Can history-based models rival globally optimized models? In *Proc. of CoNLL '11*, pages 238–246, Stroudsburg, PA, USA.