
Maximum Entropy Tiered Tagging

ALEXANDRU CEAUȘU

Research Institute for Artificial Intelligence, Romanian Academy

alceausu@racai.ro

ABSTRACT. Data sparseness in tagging highly inflectional languages with large tagsets and scarce training resources is a problem that cannot be addressed using only common tagging techniques. Tiered tagging is a two-stage technique that uses for tagging a smaller "hidden" tagset and, in the second phase, recovers the original tagset using a lexicon and a set of hand-written rules. The recovering is possible only for the words contained in the lexicon. The paper describes an experiment that shows how the maximum entropy framework can be used for tiered tagging without a hand-written set of recovery rules and which works also for unknown words.

1 Tiered Tagging

The Romanian EAGLES compliant tagset, build within the MULTTEXT-EAST initiative (Erjavec 2004), has 614 morpho-syntactic description codes (MSDtags), plus 10 punctuation tags.

Tiered tagging (Tufiș 1999; Tufiș 2000) is a two-stage technique addressing the issue of data-sparseness: (i) intermediary tagging using a reduced tagset (Ctag-set), (ii) replacing the Ctags with contextually appropriate MSD tags (called in (Tufiș 1999) MSD recovery).

The lexicon, underlying the tiered tagging approach, contains the words annotated with the MSD tags, an entry having the form: *word lemma msd*. For Romanian, this lexicon contains almost 600,000 entries. Based on the MSDtag-set lexicon a Ctag-set lexicon is automatically computed. The algorithm for Ctag-set generation and controlled information loss is described in (Tufiș 2000). This lexicon-based algorithm may produce many different Ctag-sets. The information-loss Ctag-set for Romanian consists of 92 tags, plus 10 punctuation tags. Selecting the most appropriate one was a matter of expert-introspection and required various experimental trials.

To eliminate this inconvenience, (Tufiș and Dragomirescu 2004) describe a language independent algorithm for automatic construction of the "optimal" information lossless Ctag-set. This new algorithm considers the frequency of the words in training corpora, as an additional parameter of the design procedure.

The Ctag-set is derived from the MSDtag-set by repeated generalisations by leaving out some attributes and their respective values from the original tagset specification. This procedure may be information lossless, meaning that the recovering of the left-out information

is deterministic, or may be an information-loss generalisation, meaning that the recovering process would face some ambiguities which have to be solved by using some additional knowledge resource. In (Tufiş 1999) this new resource is a set of hand-written contextual disambiguation rules. Both deterministic and the rule-based successful recovering is applicable only to the words recorded in the MSDtag-set lexicon.

The unknown words are likely to appear in any realistic application that requires tagging and they are responsible for the most annotation errors. We replaced the second phase of the tiered tagging process with a maximum entropy-based MSD recovery. In this approach, the rules for Ctag to MSD conversion are automatically learnt from the corpus and their application does not require looking-up the MSD tagset lexicon. Therefore, even the Ctags assigned to unknown words can be converted into MSD tags. If an MSD-lexicon is available, replacing the Ctags for the known words by the appropriate MSD tags is almost 100% accurate. The estimated accuracy of the Ctag to MSD for unknown words is 95.2%. Moreover, the ME model for Ctag-set - MSDtag-set may disregard the initially assigned Ctag for an unknown word and produce an unrelated MSD tag which better fits in the context. This way, some wrongly tagged unknown words may receive a correct MSD tag.

2 Maximum Entropy Framework

The maximum entropy framework is well suited for tagging since it can combine diverse forms of contextual information in a principled manner. Also, maximum entropy is one of the best tagging techniques reporting 96.43% total word accuracy and 86.23% unknown word accuracy on unseen Wall St. Journal data (Ratnaparkhi 1998).

Tagging can be re-formulated as a classification problem: the task of the classifier is to extract evidence from a linguistic "context" $b \in B$ and predict a linguistic "class" $a \in A$. The classifier will derive a conditional probability distribution p , where $p(a|b)$ is the probability of "class" a given the "context" b .

The probability model combines the evidence using weights for each predicate of the context:

$$p(a|b) = \frac{1}{Z(b)} \prod_k^{j=1} \alpha_j^{f_j(a,b)} \quad (1.1)$$

$$Z(b) = \sum_a \prod_k^{j=1} \alpha_j^{f_j(a,b)} \quad (1.2)$$

where k is the number of contextual predicates and $Z(b)$ is a normalization factor. Each contextual predicate f_j has a "weight" α_j . $p(a|b)$ represents the conditional probability of a tag a , given the context b .

	Ctag tagger	MSD tagger	Tagset converter
Wordform	x	x	
character length	x	x	x
prefix (1-2)	x	x	x
suffix (1-4)	x	x	x
upper case (all, initial)	x	x	x
is abbreviation	x	x	x
has underscore	x	x	x
has number	x	x	x
hyphen position (start, middle, end, none)	x	x	x
previous MSD features		x	x
previous MSD unigram, bigram and trigram		x	x
previous Ctag unigram and bigram	x		x
next Ctag unigram and bigram			x
end of sentence punctuation mark	x	x	x

Table 1.1: Contextual predicates.

A contextual predicate, given $f(a, b)$, may be activated for any word or tag in the context b , and must encode the information that help predicting a , such as the spelling of the current word, or the preceding unigram, bigram or trigram.

The search algorithm is a top K breadth first search that maintains, for each new word, the K highest probability tag sequence candidates.

When a lexicon is available, the tagger chooses only from the tags available for the respective word. (Ratnaparkhi 1998) reports minimal increases in performance when his maxent tagger for English uses a lexicon (0.12%). We observed that this is not the case with Romanian - the tagger accuracy is increased by 1.81%.

3 Tagging and Tagset Conversion

We developed three types of maximum entropy classifiers: (i) Ctag-tagger (Ctag-set - 102 descriptors); (ii) MSD-tagger (MSDtag-set - 624 descriptors); (iii) tagset converter (Ctag to MSD). They are based on SharpEntropy (Northedge 2005), a C# port of the MaxEnt toolkit (<http://opennlp.sourceforge.net>).

The set of contextual predicates used by each of them is detailed in table 1.1.

3.1 Ctag-tagger

The Ctag-tagger has basically the same architecture as the one from the OpenNlp Maxent package. Only the context generator of the tagger was modified in order to accommodate

Wordform	Ctag	MSD
holul	NSRY	Ncmsry
blocului	NSOY	Ncmsoy
mirosea	V3	Vmii3s
a	S	Spsa
varză	NSRN	Ncfsrn
călită	ASN	Afpfsrn
și	CR	Crssp
a	TS	Spsa
preșuri	NPN	Ncfp-n
vechi	APN	Afp-p-n
.	PERIOD	PERIOD

Table 1.2: Sample data.

features that we considered important for Romanian (like the position of the hyphen, the numbers of the characters for suffix and prefix analysis, end of sentence punctuation mark, etc.).

The tagger uses the Ctag-set (around 100 tags). In this familiar tagging scenario, the Ctag-tagger outperforms an HMM tagger with 1.5% in accuracy when tagging the "1984" Romanian corpus.

3.2 MSD-tagger

To demonstrate the need of an intermediary tagging with a reduced tagset when tagging a highly inflectional language as Romanian is, we developed a tagger that uses the MSDtag-set (624 descriptors).

The MSD-tagger has basically the same context generator as the Ctag-tagger. To improve its accuracy the MSD-tagger also uses as contextual predicates the feature description encoded in the MSD labels. For example, the features generated for the morpho-syntactical descriptor "Ncmsry" will be "N0." (PoS=noun), "N1.c" (Type=common), "N2.m" (Gender=male), "N3.s" (Number=singular), "N4.r" (Case=direct) and "N5.y" (Definiteness=yes).

The MSD-tagger accuracy outperforms a HMM MSD tagger with more than 3%.

3.3 Tagset Converter

The tagset converter maps the C-tags to MSD-tags. The classifier of the tagset converter makes use of both Ctag and MSDtag contextual predicates having thus more information than the Ctag-tagger and MSD-tagger.

From the training data the tagset converter learns a partial conversion lexicon (similar to word-form lexicon) the entries of which have the form: *word msdTag₁ ... msdTag_n*.

Name	Values
Wordform	"călită"
character length	6
prefix (1-2)	"c", "că"
suffix (1-4)	"ă", "tă", "ită", "lită"
upper case (all, initial)	false
is abbreviation	false
has underscore	false
has number	false
hyphen position (start, middle, end, none)	false
previous MSD features	"PoS=noun", "Type=common", "Gender=feminine", "Number=singular", "Case=direct", "Definiteness=no"
previous MSD unigram, bigram and trigram	"Ncfsrn", "Ncfsrn,Spsa", "Ncfsrn,Spsa,Vmii3s"
previous Ctag unigram and bigram	"NSRN", "NSRN,S"
next Ctag unigram and bigram	"CR", "CR,TS"
end of sentence punctuation mark	."

Table 1.3: Contextual predicates for the word "călită".

Unknown word accuracy without word-form lexicon	95.20%
Total word accuracy without word-form lexicon	98.66%
Total word accuracy with word-form lexicon	99.04%

Table 1.4: Tagset converter accuracy on the "1984" corpus.

It also uses an a-priori non-lexicalised resource containing the complete correspondences between Ctagset and MSD tagset of the form: $Ctag\ msdTag_1 \cdots msdTag_n$. If the mapping between the tagsets is not available, it is learned from the corpus. This additional resource allows the tagset converter to generate, with high accuracy, MSD tags even for unknown or partially known words (i.e. either missing from the learnt lexicon or learnt with an incomplete ambiguity class).

In the tables 1.2 and 1.3 is an example of how the contextual predicates are selected by the context generator of the tagset converter.

The tagset converter has an accuracy of 99.04% (1.4) when an additional lexicon is available. This performance cannot be compared to the rule-based conversion approach because in the corpus we tested there are also unknown words.

3.4 Tiered Tagger

The tiered tagger is the combination of the Ctag-tagger and the MSD-tagger.

Tagging method	Ctag-tagger	MSD-tagger	Tiered tagging
Unknown word accuracy without word-form lexicon	82.24%	78.65%	78.76%
Total word accuracy without word-form lexicon	96.81%	96.22%	96.56%
Total word accuracy with word-form lexicon	98.62%	98.45%	98.58%

Table 1.5: Accuracy on the "1984" corpus.

Because it uses more contextual information than an usual tagger (it runs on the already tagged corpus), when employed by the tiered tagger, the tagset converter can also correct tagging errors on unknown words. If a word is not in the wordform-MSD lexicon the MSD tag that the model predicts may not be among the Ctag to MSD mapping alternatives. In this case, the MSD tag the model predicted is taken into account in the K breadth first search.

4 Evaluation

For our experiments we used the CONCEDE edition (Erjavec 2004) of the parallel corpus "1984" (118025 words). We kept out 1/10 of the corpus for evaluation.

In table 1.5 are presented the evaluation results. The Ctag-tagger and MSD-tagger columns display the accuracy of the ME taggers trained with the respective tagsets annotated corpora. The tagset converter column shows the accuracy of the tagset converter. The Tiered tagging column shows the accuracy of the combination between Ctag-tagger and the tagset converter.

We were especially interested in evaluating the tagging accuracy of the unknown or partially known words, and accuracy of Ctag-MSD conversion for these words. The table 1.5 shows that the tagging accuracy is significantly better when our large word-form lexicon is used, but also it shows that the C-tag to MSD conversion is reliable even without this additional resource.

The accuracy of the tiered tagging approach is better than the one of the direct MSD tagging. The difference is higher when the domain of the evaluation corpus is different from the corpus used for training (as observed in (Tufiş 1999)). Our maximum entropy tiered tagging application can reliably handle unknown words (78.76%). At a closer inspection of the conversion "errors" we noticed that several generated MSD tags which were different from the ones in the gold standard contained more information than a lexicon can provide. The most frequent case was the specification of the gender or case attributes for invariable or unmarked adjectives. This contextually deduced information appeared as result of learning an agreement rule in Romanian: the noun and its modifier must agree in gender number and case.

Bibliography

- Tufiş, Dan and Liviu Dragomirescu (2004). Tiered Tagging Revisited. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation*, LREC'2004, ELRA, Paris, pp. 39-42.
- Tufiş, Dan (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. International Conference on Language Resources and Evaluation LREC'2000, Athens, pp. 1105-1112
- Tufiş, Dan (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nth (eds) *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*, Springer, pp. 28-33
- Erjavec, Tomaž (2004). *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004, ELRA, Paris
- Erjavec, Tomaž (2001). *Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984*. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, pp. 487-492
- Ratnaparkhi, Adwait (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania, Philadelphia
- Northedge, Richard (2005). *Maximum Entropy Modeling Using SharpEntropy*. <http://www.codeproject.com/csharp/sharpentropy.asp>
- Northedge, Richard (2005). *OpenNLP Maxent machine learning package*. <http://opennlp.sourceforge.net>