Tiered tagging in a maximum entropy framework

1 Tiered tagging

Data sparseness in tagging highly inflectional languages with scarce training resources is a problem that cannot be addressed using only common tagging techniques. The Romanian EAGLES compliant tagset, build within the MULTEXT-EAST initiative (Erjavec, 2004), has 614 morpho-syntactic description (MSD) codes, plus 10 punctuation tags.

Tiered tagging (Tufis, 1998; 2000) is a two-stage technique: (i) intermediary tagging using a reduced tagset (Ctag-set), (ii) original tagset recovery. The Ctag-set subsumes the MSD tagset with controlled information loss. The full recovering of the information from the MSD left out by the Ctag-set requires some hand-written disambiguation rules. In case the Ctag-set is obtained by information loss-less generalisations, the recovering of the omitted attribute-value pairs is strictly deterministic by an additional look-up of a wordform lexicon. In (Tufis, Dragomirescu, 2004) is described a language independent algorithm for automatic construction of the "optimal" information loss-less Ctag-set.

The (information-loss) Ctag-set for Romanian consists of 92 tags, plus 10 punctuation tags. The second processing phase uses a lexicon as a mapping between the Ctag-set and the morpho-syntactic descriptors.

The rule-based phase guarantees 100% tagset conversion accuracy only for words present in the word-form lexicon of the system and if the Ctag-set was designed by information loss-less generalizations. The word-form lexicon is a collection of entries of the form *word c-tag msd* is provided. For Romanian, this lexicon contains almost 600,000 entries. Considering the fact that unknown words are the biggest problem of tagging, we present a technique for probabilistic tagset conversion that can handle unknown words also.

2 Maximum entropy tagset conversion

The maximum entropy framework is suited for tagset conversion since it combines diverse forms of contextual information in a principled manner. Also, maximum entropy is one of the best tagging technique reporting 96.6% accuracy on unseen Wall St. Journal data (Ratnaparkhi, 1998).

The probability model can be expressed as:

$$p(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

where $p(a|b)$ represents the conditional probability of a tag a, given the context b. Each parameter α_j corresponds to a feature f_j and b_i is the context available when predicting a_i.

A feature, given (a,b), may activate on any word or tag in the context *b*, and must encode any information that might help predict *a*, such as the spelling of the current word, or the preceding bigram or trigram.

spelling	wordform
	character length
	prefix (1-2)
	suffix (1-4)
	upper state (all, initial)
	is abbreviation
	has underscore
	has number
	hyphen position (start, middle, end, none)
context	previous MSD unigram, bigram and trigram
	previous Ctag unigram and bigram
	next Ctag unigram and bigram
	end of sentence punctuation mark

Table 1. Contextual predicates

Wordform	**C-tag**	**MSD**
holul	NSRY	Ncmsry
blocului	NSOY	Ncmsoy
mirosea	V3	Vmii3s
a	S	Spsa
varză	NSRN	Ncfsrn
căliță	ASN	Afpfsrn
şi	CR	Crssp
a	TS	Spsa
preşuri	NPN	Ncfp-n
vechi	APN	Afp-p-n
.	PERIOD	PERIOD

Table 2. Sample data

spelling	wordform = "căliță"
	character length = 6
	prefix (1-2) = "c", "că"
	suffix (1-4) = "ă", "tă", "iță", "liță"
	upper state (all, initial) = "none"
	is abbreviation = false
	has underscore = false
	has number = false
	hyphen position (start, middle, end, none) = "none"
context	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 3. Contextual predicates for the word "căliță" (Table 2)

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

Our ME tagger, based on SharpEntropy (Northedge, 2005), a C# port of the MaxEnt toolkit, was trained on a corpus which is annotated in terms of both MSD tagset and Ctagset. The major extension of the tagger was to incorporate context from the hidden tagset (C-tag) when tagging with MSD.

From the training data it extracts a partial conversion lexicon (similar to word-form lexicon, but much smaller) the entries of which have the form: *word Ctag msdTag*$_1$ *… msdTag*$_n$

The tagger also uses a a-priori non-lexicalised resource containing the complete correspondences between Ctagset and MSD tagset of the form: *Ctag msdTag*$_1$ *… msdTag*$_n$. This additional resource allows the tagger 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).

3 Evaluation

For our experiments we used the CONCEDE edition (Erjavec, 2001) of the parallel corpus "1984" (118025 words). We kept 1/10 of the corpus for evaluation.

Unlike in the experiments reported in (Tufis, 1998; 2000), here we did not use the word-form lexicon. 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 below shows that the tagging accuracy significantly varies when our large word-form lexicon is not used, but also shows that the C-tag to MSD conversion is quite reliable even without this additional rerource.

	Mapping accuracy	C-tag	Msd	Tiered tagging
Unknown word accuracy without word-form lexicon	0.9520	0.8224	0.7865	0.7876
Total word accuracy without word-form lexicon	0.9866	0.9681	0.9633	0.9656
Total word accuracy with word-form lexicon	0.9904	0.9862	0.9845	0.9858

Table 4. Accuracy of the maximum entropy tagger and tagset converter for the "1984" corpus

The tiered tagging approach outperforms direct MSD tagging. Although, the tagset conversion accuracy is not as good as the lexicon-driven one, it can well handle unknown words (95.2%). At a closer inspection of the conversion "errors" we noticed that several converted MSD which were different from the ones in the gold standard contained more information than a lexicon can provide which was deduced from the context. The most frequent case was the specification of the gender or case attributes for invariable or unmarked adjectives. This over-specification appeared as a result of learning an agreement rule in Romanian: the noun and its modifier must agree in gender number and case.

References:

Dan Tufis, Liviu Dragomirescu, *Tiered Tagging Revisited*. In Proceedings of the 4th LREC Conference, Lisabona, 2004, pp. 39-42

Dan Tufiş, 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, 2000, pp. 1105-1112

Tufiş, D. "Tiered Tagging and Combined Classifiers" In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33

Tomaž Erjavec, *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris, 2004

Tomaž Erjavec, Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, (pp. 487--492), Tokyo, 2001

Adwait. Ratnaparkhi, *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1998

Richard Northedge, Maximum Entropy Modeling Using SharpEntropy, http://www.codeproject.com/csharp/sharpentropy.asp, 2005