Tiered Tagging and Combined Language Models Classifiers

Dan Tufiş

RACAI-Romanian Academy, 13, '13 Septembrie', Ro-74311, Bucharest
tufis@valhalla.racai.ro

Abstract. We address the problem of morpho-syntactic disambiguation of arbitrary texts in a highly inflectional natural language. We use a large tagset (615 tags), EAGLES and MULTEXT compliant [5]. The large tagset is internally mapped onto a reduced one (82 tags), serving statistical disambiguation, and a text disambiguated in terms of this tagset is subsequently subject to a recovery process of all the information left out from the large tagset. This two step process is called *tiered tagging*. To further improve the tagging accuracy we use a combined language models classifier, a procedure that interpolates the results of tagging the same text with several register-specific language models.

1 Introduction

One issue recurrent in the tagging literature refers to the tagset dimension vs. tagging accuracy dichotomy. In general, it is believed that the larger the tagset, the poorer the accuracy of the tagging process, although some experiments [4] show that this does not always hold provided enough training data is available and the tagset cardinality varies within reasonable limits (say 100-200 tags). However, when the target tagset gets larger (600-1000 tags or even more), the problem becomes the current tagging technology. We describe tiered tagging, a two-step process, as a possible solution for reconciling the tagging accuracy with the large number of tags in the target tagset (as many highly inflectional languages require). The two levels of the tiered tagging have two different tagsets: a reduced one, used for training and producing a language model (LM) which a proper tagging needs, and a large tagset containing the same information as the small one plus supplementary lexicon information. To further improve the tiered tagging accuracy we developed a combined language models classifier which tags the input text using different register-specific LMs and, interpolating the differences, produces a final more accurate tagged text. Although the so far experiments limit to Romanian texts, our methodology, called tiered tagging with combined language models (TT-CLAM), we believe, is not language dependent.

2 Tiered tagging

For highly inflectional languages, traditional linguistics distinguishes a large number of morpho-syntactic features and associated values. For Romanian, we

constructed a word-form lexicon [9], the items of which (427983 word-forms, 38807 lemmas) were described by a set of 615 morpho-syntactic descriptors (MSDs), EAGLES and MULTEXT compliant [5]. However, not all attributes or values present in these descriptors are distributionally sensitive or equally good contextual predictors/restrictors either. Moreover, some attribute values may depend on other attribute-values of a given wordform. Based on this set of morpho-syntactic descriptors (MSDs), and for tagging purposes, we designed a reduced tagset (RT) containing 82 tags (plus 10 punctuation tags) [8]. The reduced tagset, obtained by a trial&error process, eliminated attributes or merged attribute values which were either distributionally irrelevant or fully lexicon recoverable based on the remaining attribute values. Yet some attributes and values, although fully recoverable, were preserved in the reduced tagset, because they help disambiguate the surrounding words. The main property of this reduced tagset is what we call *recoverability*, to be described as follows.

Let MAP: RT$\rightarrow LT^m$ be a function that maps a tag from RT onto an ordered set of tags from the large tagset (LT), AMB: W$\rightarrow LT^n$, a function that maps a word onto its ambiguity class (from the lexicon) and TAG: W$\rightarrow$RT, a selector that returns for a word the tag assigned by a tagger (in a specific context). Then, recoverability (as achieved in our tagset design) means:

$$CARD(AMB(w) \bigcap MAP(TAG(w))) = \begin{cases} 1 \text{ in more than } 90\% \text{ cases} \\ \geq 2 \text{ for less than } 10\% \text{ cases} \end{cases}$$

The reduced tagset has the property that one tag assigned to a given word w can be deterministically mapped back onto the appropriate MSD in the large tagset in more than 90% of the cases. Note that although this mapping is almost deterministic, one tag may be mapped differently, depending on the context and the word it is assigned to. The underlying idea of the tiered tagging is the recoverability property of the reduced tagset. Having a training corpus annotated in terms of the reduced tagset, we can build an LM that is to be used to tag new texts. Then, thanks to the recoverability property, the tags are mapped onto MSDs (the large tagset).

For the rare cases of the mapping ambiguities (when a coarse-grained tag is not mapped onto a unique MSD but onto a list of MSDs), we use 14 very simple contextual rules (regular expressions). They specify, by means of relative offsets, the local restrictions made on the current tag assignment. Our rules inspect the left, the right or both contexts with a maximum span of 4 words. Such a rule, headed by a list representing the still there ambiguity, is a sequence of pairs (*MSD: conditions*) where *conditions* is a disjunction of regular expressions which, if applied to the surrounding tokens (defined as positive or negative offsets), returns a truth-value. If *true*, then the current token is assigned the *MSD*, otherwise the next pair is tried. If no one of the conditions returns a *true* value, the mapping ambiguity remains unsolved. This happens very rarely (for less than 1% of the whole text). For instance, the following rule considers a tag class DS corresponding to two merged MSD classes (possessive pronouns and possessive determiners/adjectives).

Ps|Ds
Ds.$\alpha\beta\gamma$: (-1 Nc$\alpha\beta\gamma$y)||(-1 Af.$\alpha\beta\gamma$y)||(-1 Mo.$\alpha\beta\gamma$y)|| (-2 Af.$\alpha\beta\gamma$n and -1 Ts)||
(-2 Nc$\alpha\beta\gamma$n and -1 Ts)||(-2 Np and -1 Ts)||(-2 D..$\alpha\beta\gamma$ and -1 Ts)
Ps.$\alpha\beta\gamma$: *true*

The rule reads as follows (α, β, γ represent shared attribute values, "." represent an "any" value):

IF any of the conditions a) to g) is true
a) previous word is a definite common noun
b) previous word is a definite adjective
c) previous word is a definite ordinal numeral
d) previous words are an indefinite adjective followed by a possesive article
e) previous words are an indefinite common noun followed by a possessive article
f) previous words are an indefinite proper noun followed by a possessive article
g) previous words are a determiner followed by a possessive article
THEN choose the determiner MSD; shared attribute values set by the context
ELSE choose the pronominal MSD; shared attribute values set by the context.

The second phase of the tiered-tagging is practically error-free, so in order to improve the overall accuracy of the output, the proper statistical tagging done at the first step has to be as accurate as possible. To this end, we developed the combined language model classifier, to be described in the next section.

3 Combined language models classifiers

In general terms, a classifier is a function that, given an input example, assigns it to one of the K classes the classifier is knowledgeable of. Recent work on combined classifier methods ([1],[2],[3], [6] etc.) has shown one effective way to speed up the process of building high quality training-corpora with a corresponding cost cut-down. The combined classifier methods in POS-tagging naturally derived from the work done on the taggers evaluation. In combining classifiers, one would certainly prefer classifiers of which errors would not coincide. The basic idea in combining classifiers is that they complement each other's decisions so that the number of errors is minimized. Of different statistical tests for checking error complementarity, we used McNemar's [2] and Brill&Wu's [1].

The combined classifiers methods [1], [2] are based on the combination of the output from different taggers trained on the same data. Such an approach considerably improves single tagger performance, and its applicability relies on the assumption that the errors made by one tagger are not a subset or superset of those of another tagger. This conjecture, which we called *error complementarity*, is supported by all the experiments we know of (e.g. [1], [2] also our own tests). The difference in taggers performance is mainly explained by the tagging methods, and, to a lesser extent, by the very linguistic nature of the training data. The linguistic relevance of the training text is not easily measurable.

The proposed methodology, even though similar to the one above at first sight, is actually different: instead of using several taggers and the same training

corpus, we use one tagger (ideally, this should be the best available) but train it on various register corpora. For the work reported here, we used a modified version of Oliver Mason's QTAG `http://www-clg.bham.ac.uk/oliver/java/qtag`). Each training session, based on comparable-size corpora, results in a register specific LM. A new text is tagged with all LMs and their outputs are combined for the final result. The *combined classifier* is based on static data structures (*credibility profiles*) constructed during tagger training on various corpora. In our experiments, none of the training corpora contained less than 110,000 hand-tagged items. The credibility profile (LM dependent) encodes, among other things, the probability of correct assignment for each tag, its confusion-probabilities and the overall accuracy of the LM. Our experiments and intensive tests and evaluations with various classifiers (simple majority voting and three types of weighted voting - out of which the one based on the credibility profiles performed the best) brought evidence for several challenging hypotheses which we believe are language independent:

• the *error-complementarity* conjecture holds true for the LMs combination. We tested this conjecture with 18 LMs combinations on various texts (about 20.000 words each) in three different registers (fiction, philosophy and journalism) and no experiment contradicted it;

• a text T_i belonging to a specific register R_i is more accurately tagged with the LM_i learnt for that register than if using any other LM_j. As a consequence, by tracking which one of the (LM-dependent) classifiers came closer to the final tag assignment, one could get strong evidence for text-type/register identification (with the traditional methods further applicable) [1];

• the combined LMs classifier method does not depend on a specific tagger. The better the tagger, the better the final results.

4 Evaluation, Availability and Conclusions

Based on George Orwell's *'1984'*, Plato's *'The Republic'* and several issues from *'România Liberă'* and *'Adevărul'* (the daily newspapers with the largest distribution in Romania), we constructed three different register training corpora (fiction, philosophy and journalism). They cover all the MSDs and more than 94% of the MSD-ambiguity-classes defined in the lexicon. The three training corpora were concatenated (the *Global* corpus) and used in the generation of another LM, to be referred in the following as LM_{Global}.

For testing, we specially hand-tagged about 60,000 words from different texts in the same registers: **Fiction**, **Philosophy** and **Newspapers** (articles extracted from newspapers others than those used for training).

Table 1 shows the results of McNemar's test on various LM combinations applied to the three test corpora. Our interest was in evaluating whether the paired

[1] According to an anonymous reviewer, the idea of register identification by seeing which LM models the text the best is also strongly supported by Beeferman, Berger, Lafferty: Statistical Methods for Text Segmentation, to appear in *Machine Learning, Special Issue on Natural Language Learning* vols.1/2/3, 1999

classifiers were likely to make similar errors on new texts in the given register. The threshold for the null hypothesis with a 0.95% confidence is $\chi^2_{0.95} = 3.84146$. Accepting the null hypothesis here implies that the two classifiers are expected to make similar mistakes on texts in the given register (e.g. Rep&News for Fiction, 1984&Rep, 1984&News News&Global for PHILOSOPHY and News&Global for NEWSPAPERS)[2]. For instance, in tagging the FICTION test corpus, the classifiers based respectively on *Rep* and *News* LMs performed equally well (or better said, equally bad) with a McNemar coefficient of 1.04. Rejecting the null hypothesis means that the two classifiers would make quite different errors and one of them is expected to make fewer mistakes than the other.

Table 1. McNemar's test for pairs of classifiers

FICTION				PHILOSOPHY				NEWSPAPERS			
LM	Rep	News	Global	LM	Rep	News	Global	LM	Rep	News	Global
1984	9.28	15.35	7.56	1984	**1.41**	**1.32**	8.14	1984	14.86	66.98	59.03
Rep	*	**1.04**	35.81	Rep	*	5.63	16.24	Rep	*	20.57	13.12
News	**1.04**	*	39.58	News	5.63	*	**1.47**	News	20.57	*	**3.46**

The results of tagging with combined LMs classifiers on the test texts (not included in the training corpora) are shown in Table 2. The classifiers based on single LMs (1984, Rep, News and Global) show a high level of correct agreement with less than 1% of wrong agreement. The bottom lines in the table display the accuracy of two combined classifiers: MAJORITY (MAJ) and CREDIBILITY (CRED). The evaluation results point out at least two important things:

Table 2. Evaluation results

LM	1984	Rep	News	Global	1984	Rep	News	Global	1984	Rep	News	Global
Test Texts	Fiction (20109w)				Philosophy (20136w)				Newspapers (20038w)			
(%) single classifiers	98.51	98.15	98.16	98.67	98.31	98.21	98.41	98.50	97.63	97.97	98.37	98.24
(%) right agreement	97.09				96.70				97.15			
(%) wrong agreement	0.59				0.83				0.72			
(%) *MAJ.* combiner	98.66				98.52				98.41			
(%) *CRED.* combiner	98.78				98.57				98.45			

a) splitting a balanced training corpus (*Global* in our case) into specialised register training corpora is worth considering: although LM_{Global} generally provides better results than a model based on a subcorpus, even the simplest combiner - *majority*, scores in most cases better;

[2] One may note that similarity is not transitive, as Rep&News on PHILOSOPHY are shown not to behave similarly, in spite of the pairs 1984&Rep and 1984&News.

b) the high level of correct agreement and the negligible percentage of false agreement allow the human expert annotator to concentrate quite safely on the cases of disagreement only. With less than 2.5% of the tagged text requiring human validation (see Table 2), the hand disambiguation of large training corpora becomes a less costly task.

The combined LMs classifier tagging system works in a client-server architecture with individual classifiers running on different machines. On a Pentium-II/300, under Linux, and with most of the programs written in Java, Perl and TCL, the individual classifiers' speed was about 15,000 words/min (most of the time being spent in accessing dictionaries). A new version of the entire system (using Oracle8^{TM} and most TCL and Perl code rewritten in C and Java) is expected to improve the speed for at least 3-4 times. If the speed factor is critical, using the single LM_{Global}-based classifier is the option of choice. The tagging system described in this paper is used in a program for automatic diacritics insertion for Romanian language texts [7] with a very high level of accuracy (more than 98.5%). The platform for tiered tagging with combined LMs classifier (containing the tokenizer, QTAG* tagger, the tag-to-MSD mapping and the combined LMs classifier) is designed as a public service on the web and, alond with the required language resources for Romanian, is free (license-based) for research purposes.

Acknowledgements: The work reported here built on the main results of the Multext-East(COP106/1995) and TELRI(COP200/1995) European projects and was partly funded by a grant of the Romanian Academy (GAR188/1998).

References

1. Brill, E., and Wu, J. (1998): Classifier Combination for Improved Lexical Disambiguation *In Proceedings of COLING-ACL98* Montreal, Canada, 191-195
2. Dietterich, T. (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, 1998, http://www.cs.orst.edu/~tgd/cv/pubs.html).
3. Dietterich, T. (1997): Machine Learning Research: Four Current Directions, *In AI Magazine*, Winter, 97-136
4. Elworthy, D. (1995): Tagset Design and Inflected Languages, *In Proceedings of the ACL SIGDAT Workshop*, Dublin, Ireland (also available as cmp-lg archive 9504002)
5. Erjavec, T., Monachini., M. eds. (1997):Specifications and Notation for Lexicon Encoding of Eastern Languages. *Deliverable 1.1F Multext-East* `http://nl.ijs.si/ME`
6. v. Halteren, H., Zavrel, J., and Daelemans, W.(1998): Improving Data Driven Wordclass Tagging by System Combination *In Proceedings of COLING-ACL98,* Montreal, Canada, 491-497
7. Tufiş, D., Chiţu, A. (1999): Automatic insertion of diacritics in Romanian Texts, *In proceedings of COMPLEX99,* Pecs, Hungary
8. Tufiş, D., Mason O. (1998): Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger *In Proceedings of First International Conference on Language Resources and Evaluation,* Granada, Spain, 589-596
9. Tufiş, D., Barbu, A. M., Pătraşcu, V., Rotariu, G., Popescu C. (1997). "Corpora and Corpus-Based Morpho-Lexical Processing" in Dan Tufiş, Poul Andersen (eds.) *Recent Advances in Romanian Language Technology,* Editura Academiei, 35-56 (also available at http://www.racai.ro/books)