Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

Silviu Cucerzan and David Yarowsky
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
{silviu,yarowsky}@cs.jhu.edu

Abstract

This paper presents a method for bootstrapping a fine-grained, broad-coverage part-of-speech (POS) tagger in a new language using only one person-day of data acquisition effort. It requires only three resources, which are currently readily available in 60-100 world languages: (1) an online or hard-copy pocket-sized bilingual dictionary, (2) a basic library reference grammar, and (3) access to an existing monolingual text corpus in the language. The algorithm begins by inducing initial lexical POS distributions from English translations in a bilingual dictionary without POS tags. It handles irregular, regular and semi-regular morphology through a robust generative model using weighted Levenshtein alignments. Unsupervised induction of grammatical gender is performed via global modeling of context-window feature agreement. Using a combination of these and other evidence sources, interactive training of context and lexical prior models are accomplished for fine-grained POS tag spaces. Experiments show high accuracy, fine-grained tag resolution with minimal new human effort.

1 Introduction

Previous work in minimally supervised language learning has defined *minimal* using several different criteria. Some have assumed only partially tagged training corpora (Merialdo, 1994), while others have begin with small tagged seed wordlists (such as Collins and Singer (1999) and Cucerzan and Yarowsky (1999) for named-entity tagging). Others have exploited the automatic transfer of some already existing annotated resource in a different medium or language (such as the translingual projection of part-of-speech tags, syntactic bracketing and inflectional morphology in Yarowsky et al. (2001), requiring no direct supervision in the foreign language). Ngai and Yarowsky (2000) observed that an often more practical measure of the degree of supervision is not simply the quantity of annotated words, but the total weighted human labor and resource costs of different modes of supervision (allowing manual rule writing to be compared directly with active learning on a common cost-performance learning curve).

In this paper we observe that another useful measure of (minimal) supervision is the additional cost of obtaining a desired functionality from existing commonly available knowledge sources. In particular, we note that for a remarkably wide range of languages, academic libraries, many booksellers and websites offer a foundation of linguistic wisdom in reference grammars and dictionaries. Thus starting from this baseline, what is the *marginal* cost of distilling from and augmenting this existing knowledge to achieve a desired new task functionality?

2 Inducing POS Tag Candidates from Unlabeled Bilingual Dictionaries

A substantial percentage of foreign language dictionaries that are available on line or in smaller paperback format are simple bilingual word or phrase translation lists which fail to specify part of speech.[1]

Thus one component question of this work is how can one extract preliminary part-of-speech distributions from untagged monolingual translation lists. Figure 1 illustrates such a bilingual dictionary, also specifying the true part of speech for each possible translation, which we do not assume to be generally available.

One approach is to take an unweighted mixture of the prior part-of-speech distributions for the English words e_i given in the translation list (TL) as illustrated in Figure 2. These probabilities may be estimated from a large and preferably balanced, corpus. In this work, we used statistics from the Brown and WSJ corpora combined.

[1] In this section, we will use the term POS tag to denote only the main part-of-speech tags (noun, verb, adjective, adverb, preposition, etc.) and not the fine-grained tags (such as `Noun-Genitive-fem-plur-def`).

	True	
Romanian	POS	English translation list
mandat	N	warrant; proxy; mandate; money order; power of attorney
manechin	N	model, dummy
manifesta	V	arise, express itself, show
manual	Adj	manual;
	N	manual; textbook; handbook
mare	Adj	large; big; great; tall; old; important;
	N	sea
maro	Adj	brown, chestnut

Figure 1: A sample Romanian-English dictionary. The POS tags are used only for evaluation and are not available in many bilingual dictionaries.

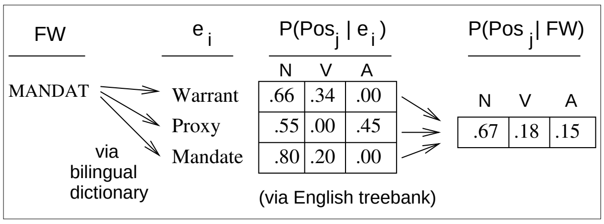

Figure 2: Inducing a preliminary POS distribution for the Romanian word *mandat* via a simple English translation list.

However, when a translation candidate is phrasal (e.g. *mandat* $\leftrightarrow$ *money order*), one can model the more general probability of the foreign word's part of speech tag (T_f) given the part of speech sequence of the English phrasal translation ($T_{e_1}...T_{e_n}$). For example, one could model $P(T_f|money\ order)$ via $P(T_f|N_eN_e)$ and $P(T_f|manifest\ itself)$ via $P(T_f|V_ePro_e)$. However, because English words often have multiple parts of speech (e.g. *order* may be a verb), one may weight phrasal POS sequence probabilities (making an independence assumption) as:

$$\begin{aligned}
&P(N_f|money\ \ order) = \\
&\quad P(N_f|N_eN_e)\cdot P(N_e|money)\cdot P(N_e|order)+ \\
&\quad P(N_f|N_eV_e)\cdot P(N_e|money)\cdot P(V_e|order)+ \\
&\quad P(N_f|V_eN_e)\cdot P(V_e|money)\cdot P(N_e|order)+ \\
&\quad P(N_f|V_eV_e)\cdot P(V_e|money)\cdot P(V_e|order)+ \\
&\quad \ldots
\end{aligned}$$

And in general:

$$P(T_f|w_{e_1}w_{e_2}) = \sum_{T_{e_1}}\sum_{T_{e_2}} P(T_f|T_{e_1}T_{e_2})\cdot P(T_{e_1}|w_{e_1})\cdot P(T_{e_2}|w_{e_2})$$

where $P(T_{e_i}|w_{e_i})$ is estimated from the dictionary as above. Without an independence assumption:

$$\begin{aligned}
&P(T_f|w_{e_1}...w_{e_n}) = \\
&\quad P(T_f|T_{e_1}...T_{e_n})\cdot P(T_{e_1}...T_{e_n}|w_{e_1}...w_{e_n})
\end{aligned}$$

There are two major options via which one can estimate $P(T_f|T_{e_1}...T_{e_n})$. The first is to assume that the part-of-speech usage of phrasal (English) translations is generally consistent across dictionaries (e.g. $P(N_f|N_{e_1}N_{e_2})$ remains high regardless of publisher or language). Hence one could use any foreign-English bilingual dictionary that also includes the true foreign word part of speech in addition to its translations to train these probabilities. Alternately, one could do a first-pass assignment of foreign-word part of speech based on only single word translations as in Figure 2, and use this to train $P(T_f|T_{e_1}...T_{e_n})$ for those foreign words having both phrasal and single-word definitions (such as *mandat*). The advantage of this approach is that it may benefit dictionaries with different phrasal translation styles from the training dictionary (e.g. use or omission of the word 'to' in verb definitions). However, given the assumption of relatively consistent dictionary formatting styles (which was unfortunately not the case for Kurdish), we evaluated this work based on supervised phrasal training from a single independent third language dictionary.

Table 1 measures the POS induction performance on three languages, where the true POS tags were given in the dictionary (as in Figure 1), but ignored except for evaluation. The accuracy values in this table are based on exact matches between a word's dictionary-provided POS and the most probable tag in its induced distribution.

For our target application of part-of-speech tagging, what matters is to have a robust tag probability distribution that includes the true candidate with sufficiently large probability to seed further training. By setting this baseline threshold to 0.1 and deleting lower ranked candidates, up to 98% of the true POS were found to be above this threshold and hence were considered in future training.

The Mean Probability of Truth, as shown in Table 1, is another measure of the quality of the POS predictions made by the algorithm, representing the probability mass associated with the true POS tag averaged over all words.

In some cases the algorithm could not predict a POS tag, primarily due to English translations for which no POS distribution was known (often an obscure word, proper name or OCR error). This oc-

Target Language	Training Dictionary	Accuracy Exact POS	Correct POS Over Threshold	Coverage	Mean Probability of Truth
Romanian	Spanish - English	92.9	97.8	98	.91
Kurdish	Spanish - English	76.8	93.1	95	.82
Spanish	Romanian - English	83.3	94.9	97	.86

Table 1: Performance of inducing candidate part-of-speech distributions derived solely from untagged English translation lists. Results are measured by type (all dictionary entries are weighted equally).

casional omission is measured by the coverage column.

Most of the observed errors are due to differences in phrasal definitional conventions in the training and testing dictionaries, long phrasal idioms, single-word definitions with ambiguous English parts-of-speech and OCR errors. The Kurdish dictionary was particularly hindered by frequent long phrasal translations which often included an explanation or definition in their translation. Because all dictionary entries are equally weighted, errors on rare words such as mythological characters or kinship terms can substantially downgrade performance. But for the purposes of providing seed POS distributions to context-sensitive taggers, performance is quite adequate for this follow-on task.

3 Inducing Morphological Analyses

There has been extensive previous work in the supervised and minimally supervised induction of both affix paradigms (e.g. Goldsmith, 2000; Snover and Brent, 2001) and diverse models of regular and irregular concatenative and non-concatenative morphology (e.g. Schone and Jurafsky, 2000; van den Bosch and Daelemans, 1999; Yarowsky and Wicentowski, 2000). While such approaches are important from the perspective of learning theory or broad coverage handling of irregular forms, another possible paradigm for minimal supervision is to begin with whatever knowledge can be efficiently manually entered from the grammar book in several hours work.

Root Affix	Inflected Affix	Part-of-speech Tag
Spanish:		
o$	o$	Adj-masc-sing
o$	os$	Adj-masc-plur
o$	a$	Adj-fem-sing
o$	as$	Adj-fem-plur
e$	e$	Adj-masc,fem-sing
e$	es$	Adj-masc,fem-plur
ar$	o$	Verb-Indic_Pres-p1-sing
ar$	as$	Verb-Indic_Pres-p2-sing
ar$	a$	Verb-Indic_Pres-p3-sing
ar$	amos$	Verb-Indic_Pres-p1-plur
ar$	áis$	Verb-Indic_Pres-p2-plur
ar$	an$	Verb-Indic_Pres-p3-plur
Romanian:		
ā$	e$	Noun-Nomin-p3-fem-plur-indef
e$	i$	Noun-Nomin-p3-fem-plur-indef
ea$	ele$	Noun-Nomin-p3-fem-plur-indef
i$	ile$	Noun-Nomin-p3-fem-plur-indef
a$	ale$	Noun-Nomin-p3-fem-plur-indef
$	$	Adj-masc,neut-sing
$	ā$	Adj-fem-sing
$	i$	Adj-masc,neut,fem-plur
$	e$	Adj-fem,neut-plur
ru$	ra$	Adj-fem-sing
ru$	ri$	Adj-masc,neut,fem-plur
ru$	re$	Adj-fem-plur
...	...	...
e$	$	Verb-Indic_Pres-p1-sing
e$	i$	Verb-Indic_Pres-p2-sing
e$	e$	Verb-Indic_Pres-p3-sing
e$	em$	Verb-Indic_Pres-p1-plur
e$	eţi$	Verb-Indic_Pres-p2-plur
e$	$	Verb-Indic_Pres-p3-plur

Table 2: Sample extracted regular inflectional paradigms (suffix context is marked by $).

We defined such grammar-based "supervision" as entry of regular inflectional affix changes and their associated part of speech in standardized ordering of fine-grained attributes, as in Table 2 for Spanish and Romanian. The full tables have approximately 200 lines each and required roughly 1.5-2 person-hours for entry.

Given a dictionary marked with core parts of speech, it is trivial to generate hypothesized inflected forms following the regular paradigms, as shown in the left size of Figure 3. However, due to irregularities and semi-regularities such as stem-changes, such generation will clearly have substantial inaccuracies and overgenerations.

However, through weighted-Levenshtein-based iterative alignment models, such as described in Yarowsky and Wicentowski (2000), one can perform a probabilistic string match from all lexical tokens actually observed in a monolingual corpus, as

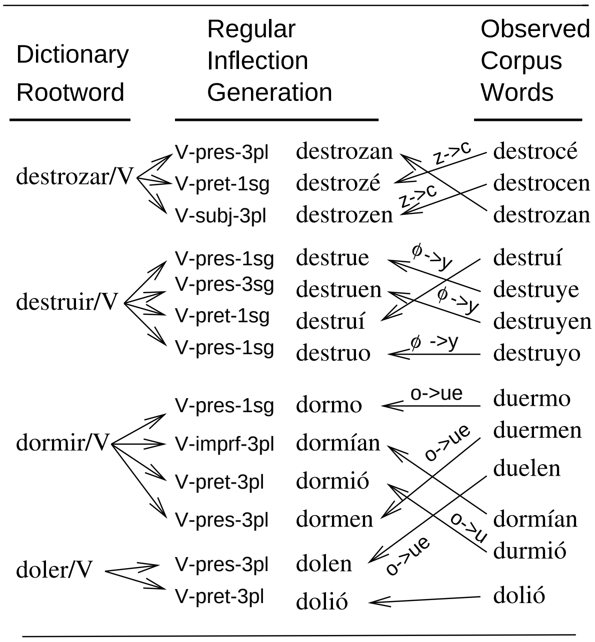

Figure 3: Inflectional analysis induction via weighted string alignment to noisy generations from dictionary roots under regular paradigms

in the right side of Figure 3[2].

For example, when looking for a potential analysis path for the Spanish irregular inflection *destrocen*, the closest string match is the regular hypothesis *destrozar/V* $\rightarrow$ *destrozen*/V-pres_subj-3pl. Likewise, the closest string match for *destruyen* is *destruir/V* $\rightarrow$ *destruen*/V-pres_indic-3pl. The differences between these regular hypotheses and observed inflected forms are the relatively productive stem changes $\emptyset \rightarrow y$ and $z \rightarrow c$, neither of which was listed in the inflectional supervision table, and yet they were correctly handled. Note that a traditional $P(\text{POS}|\text{suffix})$ model would fail to handle this case given that the common inflection suffix *-en* corresponds to two different parts of speech here (present indicative or subjunctive depending on *-ir* or *-ar* paradigm).

Also note that the irregular stem change processes such as *dormir*$\rightarrow$*duermen* have a correct best-fit analysis, despite the absence of any internal stem change exemplars (e.g. *o*$\rightarrow$*ue*) in the human-generated inflectional supervision table.

For further robustness, the consensus model of $P(Pos_i|FW)$ is estimated as a weighted mixture of the part-of-speech tags of the most closely aligned pseudo-regular generated inflections.

The inflections of closed-class words (such as pronouns, determiners and auxiliary verbs) are not well handled by this generative-alignment model, both due to their often very high irregularity (e.g. the Spanish verb *ser* (to be)) and/or their typical shortness (e.g. the pronominal inflections of *mi, tu, su*). Thus as one final amount of supervision, lists of closed-class words, paired with their inflections and fine-grained part-of-speech tags were entered manually from the grammar book (e.g. `aquellas#(aquel)Adj_Dem-fem-plur-p3`). This final source of supervision utilized an average of 400 lines and 3 person-hours per language.

4 POS Model Induction

The non-traditional supervision methodology in Sections 2 and 3 yields a noisy but broad-coverage candidate space of parts of speech with little human effort.

We then perform a noise-robust combination of model estimation and re-estimation techniques for the syntagmatic trigram models $P(pos_2|pos_1, pos_0)$ and lexical priors $P(w_i|pos_j)$ using the word co-occurrence information from a raw corpus.

- A suffix-based part-of-speech probability model $P(pos_j|\mathit{suffix}(w_i))$ using hierarchically smoothed tries is trained on the raw initial tag distributions, yielding coverage to unseen words and smoothing of low-confidence initial tag assignments.
- Paradigmatic cross-context tag modeling is performed as in Cucerzan and Yarowsky (2000) when sufficiently large unannotated corpora are available.
- Sub-part-of-speech contextual agreement for features such as gender is performed as described in Section 4.1.
- The part-of-speech tag sequence models $P(pos_2|pos_1, pos_0)$ utilize a weighted backoff between fine-grained and coarse-grained tags.
- Both the tag-sequence and lexical prior models are iteratively retrained using these additional evidence sources and first-pass probability distributions.

The success of this model is based on the assumption that (a) words of the same part of speech tend to have similar tag sequence behavior, and (b)

[2]For processing efficiency, one additional constraint is that potential hypothesized$\leftrightarrow$observed string pair candidates must exactly match in both initial consonant cluster and suffix of the generated hypothesis.

there are sufficient instances of each POS tag labeled by either the morphology models or closed-class entries described in Section 3. One example where these assumptions do not hold is for the Romanian word *a*, which has 5 possible POS tags, including `Infinitive_Marker` (corresponding to the English word *to*). But because the `Infinitive_Marker` tag has no other word instances in Romanian, no other filial supervision exists to resolve the ambiguity of *a* if no context-sensitive tagging is provided (such as the preference for *a* to be labeled `Infinitive_Marker` when followed by a `Verb-Infinitive`). Thus one avenue of potential improvement to these models would be to include limited tagged contexts for ambiguous small class (or singleton class) words, although such supervision is less readily extractable from grammar books by non-native speakers, and was not employed here.

4.1 Contextual-agreement models for part-of-speech subtags

Traditional part-of-speech models assume a strict Markovian sequential dependency. However, Adj-Noun, Det-Noun and Noun-Verb agreement at the subtag-level (e.g. for person, number, case and gender) often do not require direct adjacency, and are based on the selective matching of isolated subfeatures. This is particularly important for grammatical gender, where the lack of gender features projected from English rootwords in a bilingual dictionary (as in Section 2) require contextual agreement to assign gender to many inflected and root forms.

However, given the assumptions of minimal supervision, it is not reasonable to require a parser or dependency model to identify non-adjacent agreeing pairs explicitly. Rather, we utilize a much more general tendency for words exhibiting a property such as grammatical gender to co-occur in a relatively narrow window with other words of the same gender (etc.) with a probability greater than chance. Empirically, we observe this in Figures 4-5, which show the gender-agreement ratio between a target noun/adjective and other gender marked words appearing in context at relative position $\pm i$. Adjectives in Romanian exhibit a stronger agreement tendency with words to their left (5/1 ratio), while for nouns the agreement ratio is quite closely balanced between -1 (primarily determiners) and +1 (primarily adjectives), although weaker (2.4/1 ratio), perhaps due to a greater relative tendency for nouns to juxtapose directly with other independent clauses of different gender. Also, both parts of speech converge on the agreement ratio expected by chance (0.82) relatively quickly. Thus while any individual context may suggest incorrect gender based on agreement, if one aggregates over all occurrences of a word in a corpus, a consensus gender preference emerges, with the true gender agreement signal exceeding nearby spurious gender noise.

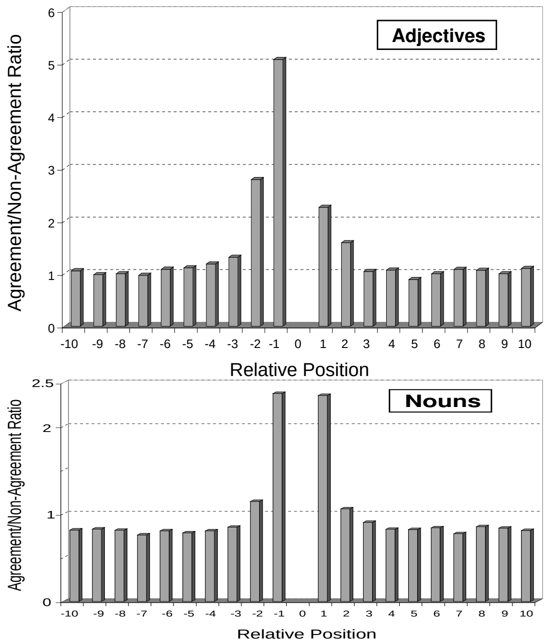

Figure 4: Ratio of the frequency that a gender-marked **adjective** (above) or **noun** (below) agrees in gender with another noun/adjective/determiner at relative position *i* over the frequency of gender disagreement at that relative position.

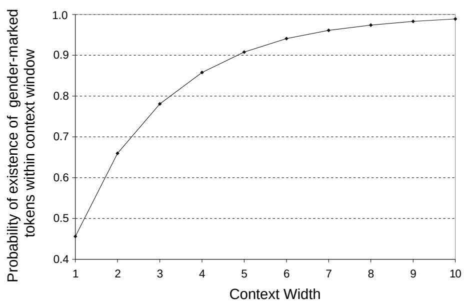

Figure 5: The probability that at least one gender-marked word will occur within a window of $\pm i$ words relative to another gender marked word (of any part of speech).

Formally, we can model this window-weighted global feature consensus as:

$$P(Gen_k|w) = \frac{1}{N} \sum_{i \in loc(w)} \sum_{j=-3}^{+3} P(Gen_k|w_{i+j}) Wt(j)$$

The ± 3 window-size parameter was selected prior to the studies shown in Figures 4-5, but is supported by them. Beyond this window the agreement/disagreement ratio approaches chance, but with a smaller window the probability of finding any gender-marked word in the window drops below the 80% coverage observed for ± 3, trading lower coverage for increased accuracy.

If one makes the assumption that the overwhelming majority of nouns have a single grammatical gender independent of context, we perform smoothing to force nouns with sufficient global context frequency towards their single most likely gender.

Finally, the trie-based suffix model noted in Section 3 can be utilized here to further generalize gender affixal tendencies for use in smoothing poorly represented single words. Through this approach we successfully discover a wide space of low-entropy gender affix tendencies, including the common *-a, -dad* and *-ción* feminine affixes in Spanish, without any human or dictionary supervision of nominal gender. But even those words without gender-distinguishing affixes (e.g. *parte*, *cabal*) can be successfully learned via global context maximization.

5 Evaluation of the Full Part-of-speech Tagger

One problem with minimally supervised learning of foreign languages is that annotated evaluation data are often not available for the features being induced, or are otherwise difficult to obtain. Thus we have used for initial test languages two languages familiar to the authors (Romanian and Spanish) for which sufficient evaluation resources could be obtained. However, the monolingual corpora utilized for bootstrapping were quite small (123 thousand words of the book *1984* for Romanian and 3.2 million words of newswire for Spanish), which are easily comparable to the sizes that can be accessed online for 60-100 world languages. The seed dictionaries were located online (for Spanish - 42k entries) and via OCR (for Romanian - 7k entries), and small grammar references were obtained at a local bookstore. 1000 words of test data were annotated with a standardized, finely detailed part-of-speech tag inventory including the full complex distinctions for gender, person, number, case, detailed tense and nominal definiteness (an inventory of 259 and 230 fine-grained tags were used for Spanish and Romanian respectively).

The minimal supervision in this study consisted of an average total of 4 person-hours per language for manually entering the inflectional paradigms and associated parts of speech from a grammar as in Section 3, and an additional average of 3 person-hours per language for dictionary extraction and entry parsing. OCR itself on our high-speed 2-sided scanner with OmniPage Pro took under 30 minutes). As would be expected given that data entry was done by computer scientists which were not native speakers of the test languages, significant analysis errors or gaps were introduced when rather blindly transferring from the reference grammar. Thus to test the relative contributions of limited native speaker help when available, for roughly 4 additional total person hours in a second test condition for Romanian a native speaker corrected and augmented gaps in the patterns previously entered from the grammar book, focusing almost exclusively on the complex inflections of closed-class words.

A summary of the results for these three supervision modes is given in Table 3. Performance is broken down by fine-grained part of speech. Exact-match accuracy is measured over both the full fine-grained (up to 5-feature) part-of-speech space, as well as the 12-class core POS tag (noun and proper noun, pronoun, verb, adjective, adverb, numeral, determiner, conjunction, preposition, interjection, particle, punctuation). The feature of grammatical gender was specifically isolated because it is rarely salient for cross-language applications such as machine translation (where grammatical gender rarely transfers), and because its induction algorithm in Section 4.1 depends heavily on the size of the monolingual corpus (which is small in these experiments, suggesting size-dependent potential for significant further improvement here).

Finally, a post-hoc analysis of the system vs. test data discrepancies showed that a significant number were simply arbitrary differences in annotation convention between the grammar-book analyses and the test data tagging policy. For example, one such "error"/discrepancy is the rather arbitrary distinction of whether the Romanian word *oricare* (meaning *any*) should be considered an adjective (as listed in a standard bilingual dictionary) or a determiner. Another difference is whether proper-name citations of common nouns (e.g. *Casa Blanca*) should be annotated for gender/number etc. or not.

Yet regardless of exactly how many system-test discrepancies are just policy differences rather than errors, even the raw accuracy here is very promising given the very fined-grained part-of-speech inventory and small monolingual data size used for bootstrapping. And ultimately the performance is quite

	Spanish	Romanian	
	NNS 8h	NNS 8h	NNS-8h NS-4h
All words			
core-tag	93.1	86.3	89.2
exact-match	86.5	68.6	75.5
exact w/o gender	87.0	76.7	83.0
Nouns			
core-tag	90.3	97.4	97.4
*number	100.0	97.4	98.9
*gender	100.0	54.9	64.7
*definiteness	–	96.6	93.7
*case	–	97.4	97.4
Verbs			
core-tag	94.7	87.9	89.5
*tense	93.0	92.6	93.2
*number	100.0	91.5	91.2
*person	97.2	92.6	93.2
Adjectives			
core-tag	79.7	78.6	81.5
*gender	100.0	81.3	82.2
*number	100.0	98.3	98.3

Table 3: Performance of POS tagger induction based on 1 person-day of supervision, no tagged training corpora and a fine-grained (≈250 tags) tagset. NNS and NN refer to non-native-speaker and native-speaker effort.

remarkable given that it is the result of less than 1 total person day of data collection and supervision, in contrast to the thousands of hours and $100,000-$1,000,000 spent on some annotated training data in a much more limited tagset inventories. Thus in terms of cost-benefit analysis, the supervision paradigm and associated bootstrapping models presented here offer quite a good value of new functionality per labor invested.

6 Conclusion

This paper has presented an alternative to traditional corpus annotation-based supervision of part-of-speech taggers. Given that even obscure languages have reference grammars and dictionaries available in large bookstores, libraries or even on-line, the focus of this work is on using human supervision for efficient structured entry of this seed knowledge (in the form of regular and semi-regular inflectional paradigms and often irregular closed-class part-of-speech entries). Minimally supervised bootstrapping procedures then used corpus-derived distributional data to induce lexical tag probabilities from dictionaries, irregular morphological analyses via weighted Levenshtein-based alignment models, tag sequence probability induction and grammatical gender agreement modeling. Experiments show high accuracy coarse and fine-grained (≈ 250 tag) part-of-speech analyses using only one person day of new human supervision based on readily available linguistic resources.

Acknowledgements

This work was partially supported by NSF grant IIS-9985033 and ONR/MURI contract N00014-01-1-0685.

References

Baum, L. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Collins, M., and Y. Singer, 1999 Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999*, pp. 100-110.

Cucerzan, S., and D. Yarowsky, 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999*, pp. 90-99.

Cucerzan, S., and D. Yarowsky, 2000. Language independent minimally supervised induction of lexical probabilities. In *Proceedings of ACL 2000*, pp. 270-277.

Goldsmith, J. A., 2000 Unsupervised learning of the morphology of a natural language. *Computational Linguistics 27:2*, pp. 153-198.

Merialdo, B., 1994. Tagging English text with a probabilistic model. *Computational Linguistics 20*, pp. 155-171.

Ngai, G., and D. Yarowsky, 2000. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL 2000*, pp. 200-207.

Schone, P., and D. Jurafsky, 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of CoNLL 2000*.

Snover, M. G., and M. R. Brent, 2001. A Bayesian model for morpheme and paradigm identification. In *Proceedings of ACL 2001*, pp. 482-490.

Van den Bosch, A., and W. Daelemans, 1999. Memory-based morphological analysis. In *Proceedings of ACL 1999*, pp.285-292

Yarowsky, D., G. Ngai, and R. Wicentowski, 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001*, pp. 161-168.

Yarowsky, D., and R. Wicentowski, 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL 2000*, pp. 207-216.