Evaluating Distributional Properties of Tagsets

Markus Dickinson, Charles Jochim

Indiana University, University of Stuttgart
md7@indiana.edu, charles.jochim@ims.uni-stuttgart.de

Abstract

We investigate which distributional properties should be present in a tagset by examining different mappings of various current part-of-speech tagsets, looking at English, German, and Italian corpora. Given the importance of distributional information, we present a simple model for evaluating how a tagset mapping captures distribution, specifically by utilizing a notion of *frames* to capture the local context. In addition to an accuracy metric capturing the internal quality of a tagset, we introduce a way to evaluate the external quality of tagset mappings so that we can ensure that the mapping retains linguistically important information from the original tagset. Although most of the mappings we evaluate are motivated by linguistic concerns, we also explore an automatic, bottom-up way to define mappings, to illustrate that better distributional mappings are possible. Comparing our initial evaluations to POS tagging results, we find that more distributional tagsets can sometimes result in worse accuracy, underscring the need to carefully define the properties of a tagset.

1. Introduction

To evaluate part-of-speech (POS), or category, induction, POS tags are often mapped to a smaller tagset (e.g., Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), but there have been few criteria for evaluating the quality of these mappings. It is not always clear which properties of the tagset are more or less important to evaluate. Tagset mappings assumedly capture distributional properties, but this is not always made explicit, nor have there been many comparisons made between mappings (cf., though, Dickinson and Jochim, 2009). Given that a major weakness with evaluating category induction is that “there is not a unique well-defined set of part-of-speech tags” (Clark, 2003), we can improve the understanding of induction evaluation by better defining the properties which should be made across tagsets and by elucidating where desired mappings cannot be made.

Relatedly, a more general question is what properties should be in a tagset to begin with. One property commonly addressed is that of distribution (dating back to Harris, 1954). However, tagsets in existence tend not to cleanly delineate morphological and distributional evidence in defining a tag (though, see recent work on learner language for such a proposal, e.g., Díaz-Negrillo et al., 2010, to appear; Dickinson and Ragheb, 2009). Thus, when evaluating POS taggers, for example, it is not always clear which parts of errors stem from a lack of distributional evidence or a lack of morphological evidence, or some combination thereof. We account for this issue by comparing different tagset mappings with distinct distributional properties.

Regardless of tagset mappings, tagset evaluation is important in itself. Tagsets can be evaluated with respect to their *internal* quality, i.e., whether they can be tagged accurately, or their *external* quality, i.e., whether they make important linguistic distinctions (see Déjean, 2000, sec. 2 & 7). Typical accuracy measures reported for tasks such as POS tagging reflect internal quality, but, at least for tagset mappings, there should be ways to measure the external quality, too. Furthermore, it has been argued that it is the external properties which are of more importance in tagger evaluation (Elworthy, 1995). We address this in our paper by using a lexicon-based evaluation, in addition to accuracy measures.

In this paper, we investigate the question of which distributional properties should be in a tagset and examine the properties currently found in POS tagsets, by examining different mappings. Outlining which categories are more or less distributional can help drive the design and refinement of tagsets. Furthermore, by making external and internal criteria clearer, we argue that an evaluation metric which reflects the loss of linguistically (externally) important information in a mapping should be used.

2. Tagset mappings

Corpus POS categories are composed of a variety of morphological and syntactic features, the exact nature of which varies across tagsets. Tagset mappings provide a way to examine the linguistic properties which are appropriate for certain types of evaluation, ignoring the properties which are not. For example, by merging certain tags, we can factor out morphological properties to determine which syntactic properties are correct. For learner language, it has been proposed that POS annotation should be split into distributional, morphological, and lexicon tags, in order to capture the ways that second language learners use language non-natively (see Díaz-Negrillo et al., 2010, to appear). Such an explicit delineation of different sources of POS evidence would be useful in our context, but since tagsets do not currently have this tripartite structure, we define mappings to get at the specific distributional properties.

Investigating tagset mappings is nothing new. Previously, however, tagset mappings have been used to show the effect on POS tagging (Brants, 1997) or have focused on defining mappings *between* tagsets (Zeman, 2008; Pîrvan and Tufiş, 2006), as opposed to reducing the size of a single tagset. We focus on the situation where a tagset is being reduced in size in order to evaluate system performance, often used for unsupervised category induction, e.g., the PTB-17 tagset for English (Smith and Eisner, 2005; Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008).

2.1. Measuring tag quality

We need some way(s) to evaluate different tagset mappings, specifically measuring the quality of the mappings with respect to their distributional properties. We outline two ways to evaluate a tagset mapping, one which is based more on internal criteria (section 2.1.1.) and one which is based more on external criteria (section 2.1.2.).

2.1.1. Frame-based approach

Following Dickinson and Jochim (2009), we will use *frames* (Mintz, 2006) to test the quality of distributional mappings. The idea is relatively simple: local context, in the form of a (frequent) *frame* of two words surrounding a target word, leads to the target's categorization. For example, the frame *you _ it* generally predicts a verbal category for the target. Mintz (2002) shows this local context, in the form of a frame, leads to the target's categorization in adults, and Mintz (2003) shows that frequent frames supply category information in child language corpora. Frames also seem to be a viable way to investigate categorization of corpus data (Dickinson and Jochim, 2008, 2009).

Consider, though, that verbal tags often distinguish past tense from present tense, largely a morphological property in English. The apparent accuracy of a frame at identifying a corpus category thus suffers because non-distributional properties may be encoded in the tags. A tagset mapping which merges these properties is still largely capturing distribution.

Frames are essentially a very primitive classifier, identifying all words in the same frame as having the same category. They have a distinct advantage over more complex classifiers (e.g., POS taggers), however, in that they are one way of encoding a purely distributional test. Thus, tagset mappings which do a better job of increasing the accuracy of frame-based identification are better capturing distributional properties (see also Dickinson and Jochim, 2009). Frames are additionally beneficial, in that they are quite simple to encode and appear to be somewhat cross-linguistically viable, as has been shown for human category acquisition (Chemla et al., 2009; Xiao et al., 2006).

Defining frequency For category acquisition, *frequent frames* are used, those with a frequency above a certain threshold. For frequent frames in six subcorpora of the CHILDES database (MacWhinney, 2000), Mintz (2003) obtains both high type and token accuracy in categorizing words.

The core idea of using frames is that words used in the same context are associated with each other, and the more often these contexts occur, the more confidence we have that the frame indicates a category. Setting a threshold to obtain the 45 most frequent frames in each subcorpus (about 80,000 words on average), Mintz (2003) allows a frame to occur often enough to be meaningful and have a variety of target words in the frame.

On a par with obtaining around 45 frequent frames and following Dickinson and Jochim (2009), we define frequent frames as those which are about 0.03% of the total number of frames; this works out to be those frames with a frequency of 200 in the Penn Treebank corpus. With this cutoff, only frames that occur frequently will be categories, increasing reliability of the data. One could explore more thresholds, but for comparing tagset mappings, these provide a good picture.

Accuracy To evaluate, we need a measure of the accuracy of each frame. Mintz (2003) and Redington et al. (1998) calculate accuracy by counting all pairs of words (types or tokens) that are from the same category, divided by all possible pairs of words in a grouping. This captures the idea that each word should have the same category as every other word in its category set.

However, this measurement does not seem to adequately represent cases with a majority label. For example, if three words have the tag X and one Y, pairwise comparison results in an accuracy of 50%, even though X is dominant. To account for this, we measure the *purity* (see Ch. 16 in Manning et al., 2008) of the frame by dividing the most frequent category instances among all instances, e.g., 75% for the above example. We will use this measure throughout the paper, although, for comparison of tagset mappings on the same data, either measure seems adequate.

2.1.2. Lexicon evaluation

Measuring distributional accuracy does not capture all the relevant facts for evaluating tagset mappings. Accuracy only measures internal qualities of the tagset. We also want some way to measure the loss in external quality, i.e., the types of linguistic distinctions which are no longer made by a mapping.

To motivate our measurement, consider the fact that some category distinctions are less important for the context to make. For example, using the Penn Treebank tags for English (Marcus et al., 1993), it is detrimental if we conflate base form verb (VB) and present tense verb (non-third person singular, VBP) because this is a prominent ambiguity for many words (e.g., *see*). On the other hand, there are no words which can be both VBP (e.g., *see*) and VBZ (third person singular present tense, e.g., *sees*). Following Dickinson and Jochim (2009), we thus also report how many ambiguities are lost in the lexicon for a given tagset mapping, in order to measure the loss in desired linguistic properties. For example, mapping VB and VBP to a single verbal tag causes many words to lose ambiguities.

A mapping is preferred which does not conflate tags that vary for individual words. To calculate this, we compare the original lexicon with a mapped lexicon and count the number of words which lose a distinction. Consider the words *accept* and *accepts*: *accept* varies between VB and VBP; *accepts* is only VBZ. When we map tags based on similar form (see section 4.1.), we count 1 for *accept*, since VB and VBP are merged into one tag (Verb). When we map verbs based on finiteness, we count 0 for these two words, as *accept* still has two tags (V-nonfin, V-fin) and *accepts* has one tag (V-fin), even if they are reduced tags.

Fewer losses are desired, as this means that words are nearly as ambiguous as they were before. In the limit, the best methods will merge nothing, and the worst will merge everything into one tag, losing every ambiguity in the lexicon.

3. Initial mappings

We first look at convenient mappings for different languages, many of them currently in use for evaluation. Table 1 shows the accuracy of frequent frames using the corpora's original and mapped tagsets. We report the number of frames used for comparison, number of tags, purity, and the loss in lexicon ambiguities (*Lost amb.*). The first row in each section is the unmapped version, and subsequent rows are mappings. The mappings—which are more fully described below—get us closer to the results for human category acquisition (Mintz, 2006), but the loss of lexical ambiguity often skyrockets. And, as shown below, a lower number of mapped tags is not always correlated with higher accuracy. We will need to more carefully consider the distributional dimensions of the mappings and the tagsets (see section 4.).

Corpus mapping	Frames	Tags	Purity	Lost amb.
PTB	98	45	79.5%	0
PTB-17	98	17	89.7%	2038
Bro.	88	383	66.3%	0
Bro.-17	88	18	84.0%	580
SUS.	102	425	38.1%	0
SUS.-1	102	20	79.1%	652
SUS.-2	102	61	75.4%	589
TIG.	58	155	82.3%	0
TIG.-1	58	14	90.5%	2627
TUT	149	924	63.5%	0
TUT-1	149	16	89.6%	183
TUT-2	149	94	84.2%	64

Table 1: Original & (coarsely) mapped tag purity

PTB & Brown The PTB-17 mapping (Smith and Eisner, 2005) for the Penn Treebank is commonly used for evaluating category induction (e.g., Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008). We use a similar mapping for the Brown tagset, as they share many tags in common. The original Brown data has been preprocessed so that tokenization would be more similar to the Penn Treebank, and function tags are ignored.

SUSANNE For the SUSANNE Corpus tagset (Sampson, 1995), the tags are composed in such a way that each character makes a finer distinction. This makes it easy to map tags, by taking the first character of the tag (*SUS.-1*) or the first two characters (*SUS.-2*), as in Brants (1997).

TIGER With the TIGER corpus of German (Brants et al., 2002), we tested a mapping that results in about as many tags as the PTB-17. TIGER's accuracy is greater than the accuracies of the other corpora, both before (*TIG.*) and after mapping (*TIG.-1*).

TUT The tagset for the Turin University Treebank (TUT) of Italian (Bosco et al., 2000) is composed of syntactic categories (e.g., noun), syntactic subcategories (e.g., gender), and syntactic features, leading to a much larger original tagset. A mapped tagset can be obtained by looking just at the lexical categories and select subcategories, much like with the SUS.-1 mapping (*TUT-1*), or by following the mapping in Chanev (2005) (*TUT-2*).

4. Manipulating different properties

In this section we test linguistically-motivated mappings to find which distributional properties of a tagset can easily be mapped and what the resulting accuracy is for different mappings, using frequent frames as the basis for distributional classification. Given the predominance of verbs and nouns, we focus on linguistic properties within these categories, leaving other categories unmapped. Accuracy will of course improve by merging tags, thereby removing distinctions; what is important is for which mappings it improves most and still retains desired ambiguities in the lexicon.

4.1. Penn Treebank (PTB)

Due to its popularity in training and tagging POS tagging technology, we use the PTB as a starting point.[1] For the PTB, we merge nouns and verbs along two dimensions: their common syntactic/distributional properties or their common morphological properties (Dickinson and Jochim, 2009). We merge nouns by *noun type* or by *noun form*, as shown in table 2. Specifically, the *noun type* mapping has three noun tags: PRP [pronoun], NN/NNS [common noun], NNP/NNPS [proper noun]; and the *noun form* mapping has a different set of three tags, based on grammatical number: PRP [pronoun], NN/NNP [singular noun], NNS/NNPS [plural noun]). As shown in table 3, we merge verbs

[1] For the PTB, we used only sections 0-18 to develop the method; for all other corpora, we used the entire corpus.

either by *finiteness* (MD [modal], VBP/VBZ/VBD [finite verb], VB/VBG/VBN [nonfinite verb]) or by *verb form* (MD [modal], VB/VBP [base], VBD/VBN [-*ed*], VBG [-*ing*], VBZ [-*s*]). In the latter case, verbs with consistently similar forms are grouped—e.g., *see* can be a baseform (VB) or a present tense verb (VBP).

Noun type		Noun form	
Tag	Mapping	Tag	Mapping
PRP	pronoun	PRP	pronoun
NN	common	NN	singular
NNS	noun	NNP	noun
NNP	proper	NNS	plural
NNPS	noun	NNPS	noun

Table 2: Noun mappings for Penn Treebank

Finiteness		Verb form	
Tag	Mapping	Tag	Mapping
MD	modal	MD	modal
VBP	finite	VB	base
VBZ	verb	VBP	verb
VBD		VBZ	*-s*
VB	nonfinite	VBD	*-ed*
VBG	verb	VBN	
VBN		VBG	*-ing*

Table 3: Verb mappings for Penn Treebank

These are not the only possible mappings, of course; however, we are limited in what we can map by what is provided in the original tagset. For example, we cannot map based on verb transitivity (see section 4.3.), as this is not encoded in the PTB tagset.

Mapping	Tags	Purity	Lost amb.
PTB-17	17	89.7%	2038
N. form/V. form	41	83.2%	2653
N. type/V. form	41	84.3%	2101
N. form/Finite	39	85.1%	905
N. type/Finite	39	**86.3%**	**352**
No mappings	45	79.5%	0

Table 4: Results for Penn Treebank

In table 4, we find that merging verbs by finiteness and nouns by noun type results in high precision.[2] Using frames as a distributional test, this confirms that noun type better captures distribution than noun form. This is not to say that *noun form* is not relevant distributionally—clearly, singular and plural nouns differ with respect to distributional properties such as verbal agreement (cf. *he is* vs. *they are*). However, it seems that noun type more often is needed distributionally.

In addition to being more distributional, a mapping such as the noun type/verb finiteness one also better maintains distinctions in the lexicon. The column *Lost amb.* in table 4 shows that mapping by *noun type* and *finiteness* has fewer lost ambiguities than the other noun/verb mappings. By contrast, the PTB-17 mapping (Smith and Eisner, 2005), commonly used for evaluating category induction (Goldberg et al., 2008; Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), has a better accuracy, yet it loses distinctions for 2038 words. This lexical ambiguity measure is thus crucial for a full evaluation. Further, a mapping such as the one we propose seems better suited for evaluating distributional methods of category induction.

4.2. Brown Corpus

The Brown corpus mappings are given in tables 5 and 6. Although the base categories are different, the mapping results in essentially equivalent categories to the PTB mappings. Likewise, the results for the Brown corpus (table 7) follow the same trends as PTB. In particular, here we see that the noun type distinction boosts our accuracy while mapping by verb finiteness minimizes the lexical ambiguity (with respect to the noun form and verb form distinctions respectively). We can also note that each mapping has essentially the same number of categories, indicating the importance of examining a mapping's linguistic properties (cf. also Elworthy, 1995).

Noun type		Noun form	
Tag	Mapping	Tag	Mapping
pn, ppl ppls, ppo pps, ppss	pronoun	pn, ppl ppls, ppo pps, ppss	pronoun
nn, nns nr, nrs	common noun	nn, np nr	singular noun
np, nps	proper noun	nns, nps nrs	plural noun

Table 5: Noun mappings for Brown

We should note that this data ignores the -hl and -tl tag suffixes, representing headlines and titles (see discussion in Dickinson and Jochim, 2008). Furthermore, we made some slight changes in the the tokenization from the original Brown corpus to more closely match the PTB: contractions including apostrophes are split (to avoid compound Brown tags, i.e. pps+hvz); and possessives are split, introducing a PTB-like POS tag.

4.3. SUSANNE

Extending these noun/verb mappings to the SUSANNE corpus is more problematic, where "word

[2] The results are similar to those in Dickinson and Jochim (2009), but in that paper, we also mapped other categories beyond verbs and nouns.

Finiteness		Verb form	
Tag	Mapping	Tag	Mapping
md, md*	modal	md, md*	modal
bed, bed*, bedz, bedz*, bem, bem*, ber, ber*, bez, bez*, dod, dod*, doz, doz*, hvd, hvd*, hvz, hvz*, vbd, vbz	finite verb	be, bem, bem*, ber, ber*, do, do*, hv, hv*, vb	base verb
		bez, bez*, doz, doz*, hvz, hvz*, vbz	*-s*
		beg, hvg, vbg	*-ing*
be, beg, ben, do, do*, hv, hv*, hvg, hvn, vb, vbg, vbn	nonfinite verb	bed, bed*, bedz, bedz*, ben, dod, dod*, hvd, hvd*, hvn, vbd, vbn	*-ed*

Table 6: Verb mappings for Brown

Mapping	Tags	Purity	Lost amb.
Bro.-17	18	84.0%	580
N. form/V. form	59	72.0%	1685
N. type/V. form	58	79.1%	1611
N. form/Finite	57	73.4%	188
N. type/Finite	**56**	**80.5%**	**114**
No mappings	383	66.3%	0

Table 7: Results for Brown

classes" are more lexically motivated. The mappings are given in tables 8 and 9,[3] but they are not exact. Regarding noun form mappings, for example, SUSANNE has tags for singular nouns (e.g. NN1c), plural nouns (e.g. NN2), and nouns which can be both singular and plural (e.g. NNc). This means that instances of words like *sheep* are not disambiguated; they are tagged NNc for any number. We group them with singular nouns, though this is not ideal. In other words, despite the larger tagset and finer granularity, the tagset does not help as much in our efforts of creating mappings to isolate distributional tagset properties.

Despite this limitation, the trends still parallel those of PTB and Brown, as we report in table 10. Mapping verb finiteness better retains lexical ambiguities, while mapping noun type seems to help accuracy. Because information on verbs' transitivity is present in SUSANNE,[4] we also tried such mappings: this performs worse than verb form or finiteness, indicating that it may be a more difficult distributional property.

Noun type		Noun form	
Tag	Mapping	Tag	Mapping
PP	pronoun	PP	pronoun
NP	proper noun	N..2 NNmm	plural noun
(not NP)	common noun	(not plural)	singular noun

Table 8: Noun mappings for SUSANNE

Finiteness		Verb form	
Tag	Mapping	Tag	Mapping
VM	modal	VM	modal
V.D V.Z V.M V.R	finite verb	V.0 V.M V.R	base verb
		V.Z	*-s*
V.0 V.G V.N	nonfinite verb	V.D V.N	*-ed*
		V.G	*-ing*

Table 9: Verb mappings for SUSANNE

Mapping	Tags	Purity	Lost amb.
First letter	20	79.1%	652
Two letters	61	75.4%	589
N. form/V. form	279	67.3%	532
N. type/V. form	279	73.9%	533
N. form/Finite	277	68.4%	104
N. type/Finite	**277**	**75.0%**	**105**
N. form/Trans.	279	62.9%	530
N. type/Trans.	279	69.4%	531
No mappings	425	38.1%	0

Table 10: Results for SUSANNE

4.4. Turin University Treebank (TUT)

Table 11 provides some mappings to verbs in the TUT corpus, but we do not list out the mappings for noun type, noun form, or verb form. Noun type is simple: NOUN COMMON maps to *common noun* and NOUN PROPER maps to *proper noun*. On the other hand, because Italian has more complex morphology than English, noun and verb inflections cannot be easily mapped by form. Our decision is to approximate this by mapping nouns to a gender/number class (e.g., NOUN F SING) and by mapping verbs to a person/number class (e.g., VERB 3 SING), in addition to including categories for gerunds, infinitives, modals,

[3] *not NP* means that all noun tags other than NP are merged; the period (.) indicates any possibility; tags with more than 3 characters are subsumed by the 3-character tags.

[4] These are easily mapped, given that transitive verb tags end in *t* and intransitive in *i*.

and participles. Similar to SUSANNE, features such as "ALLVAL" which indicate that all values could apply pose a problem for distributional definitions. In these cases, we mapped ALLVAL to itself.

The results (table 12) still list noun form and verb form, which shows that fewer tags are mapped.[5] We again see an accuracy increase for the noun type distinction (around 9%), but here the lexical ambiguity increases, as well, albeit slightly. Interestingly, all verb mappings have relatively similar accuracies, perhaps indicating the importance of distribution for making all the verbal distinctions in Italian. The noun type and finite mappings proposed for the English Penn Treebank still make sense for Italian, but perhaps there are additional mappings that might be useful as well.

Finiteness	
Tag	Mapping
VERB MOD	modal
VERB . IND VERB . CONG VERB . CONDIZ VERB . IMPER	finite verb
VERB . GERUND VERB . INFINITE VERB . PARTICIPLE	nonfinite verb

Table 11: Verb mappings for TUT

Mapping	Tags	Purity	Lost amb.
"syntactic categories"	16	89.6%	183
Chanev mapping	94	84.2%	64
N. form/V. form	284	75.7%	62
N. type/V. form	277	84.5%	71
N. form/Finite	269	77.1%	63
N. type/Finite	**262**	**85.8%**	**72**
N. form/Trans.	270	75.6%	57
N. type/Trans.	263	84.5%	66
No mappings	924	63.5%	0

Table 12: Results for TUT

5. Automatically mapping similar tags

The approaches for tagset mapping in previous sections were more top-down approaches: we started with a set of grammatical categories and assigned mappings for sets of categories, based on what we suspected were useful properties. With large tagsets, such as the SUSANNE tagset, this sort of approach can sometimes be time-consuming. It would be useful to have some automatic help in defining a mapping.

To that end, we use a cosine similarity measure to tell us which tags appear in the same frame contexts. Tags with similar distributions are grouped together to define a mapping. Although this is "cheating" by using the same data for measuring and evaluating, we do this in order to define a mapping which can be used for evaluation purposes. Additionally, we will see that this methodology confirms the fact that better mappings than the ones we hand-created can be defined.

Cosine similarity provides a bottom-up approach so that we can group tags based strictly on distributional properties (i.e. frames). Similar tags, where $sim(tag_1, tag_2) > 0.75$, are transitively grouped together and these groups form the mappings used. For example, calculating over the 102 frames for SUSANNE, RRQr and VV0t have 0.79 similarity, and VV0t and VV0v have 0.96 similarity. Regardless of the similarity between RRQr and VV0v, we merge these three tags into a mapping. With this methodology for the SUSANNE corpus, we obtain increased accuracy (73.3%), while keeping the lost ambiguities to a minimum (36), as shown in table 13, where previous results are included for comparison. This is despite having more tags (326) than with any other mapping.

Mapping	Tags	Purity	Lost amb.
First letter	20	79.1%	652
Two letters	61	75.4%	589
N. type/Finite	277	75.0%	105
Cosine sim.	326	73.3%	36
No mappings	425	38.1%	0

Table 13: Cosine similarity results for SUSANNE

Similarly, trying this on Brown, we obtain 78.4% accuracy with only 68 lost ambiguities, as shown in table 14. With this bottom-up approach, accuracy improves and lost ambiguities remain low; that is, this similarity-based mapping results in a good balance of internal and external criteria. However, it is not clear what the linguistic consequences are. In the future, one can consider applying linguistic intuition from the top to improve the automatic mappings of similar tags. What we have demonstrated are: 1) better mappings do exist than ones gleaned solely from linguistic intuition, and 2) once again, the number of tags in a tagset (mapping) is not correlated with either internal or external accuracy.

Mapping	Tags	Purity	Lost amb.
Bro.-17	18	84.0%	580
N. type/Finite	56	80.5%	114
Cosine sim.	79	78.4%	68
No mappings	383	66.3%	0

Table 14: Cosine similarity results for Brown

[5] Dates and numbers are mapped as DATE and NUM in all mappings, (see Chanev, 2005, for details).

6. POS tagging evaluation

6.1. Tagger-based approach

Given this analysis of different tagset mappings, we can now investigate what happens when we run a supervised POS tagger and map the results in different ways. POS tagging provides a more sophisticated method of looking at a tag distribution, but is crucially different from using frames as a classification method. First, tagging is supervised, while frames are not, meaning that a tagger reduces its set of distributional choices to those which are consistent with its lexicon. To alleviate this to some extent, we focus on the tagging accuracy of unknown words in the testing data. However, taggers still group together statistics of known words to tag unknown words. That is to say, upon seeing an unknown word, a tagger is predisposed to tag it a certain way, based on the lexicon, as opposed to using only distributional information, as with frequent frames.

Secondly, using frames to classify requires only examining frequent contexts; this limited set of contexts contrasts with a tagger that tags every word. When examining rarer contexts, the distributional evidence is less clear, i.e., there are many situations the tagger will have never seen before. It is clearly desirable to have a tagger tag these cases, but it makes a POS tagger non-optimal for evaluating distributional properties of tagsets. If we cannot reliably classify each context as fitting a particular POS tag, then there is much more guessing taking place.

Relatedly, although both POS tagging and frame-based classification capture distribution, they have a different distributional model. Frames focus on one word before and one word after a target word, whereas POS taggers are generally concerned with the context of the previous tags before the target word (at least the Markov model tagger we use).

6.2. Results

We used the TnT POS tagger (Brants, 2000) on the PTB data for our experiments. By default, TnT uses a suffix trie to guess the tags of unknown words using morphological information. We turn off the use of this trie to make the tagger rely more purely on distribution (i.e., the suffix trie is set to zero length).

It should also be noted that two very different things are measured for frame-based classification and POS tagging. For frames, we are measuring the uniformity of a context, while POS tagging accuracy refers to the accuracy of the trained tagger tested against a gold standard.

Results for POS tagging are given in tables 15 (all words) and 16 (unknown words). We focus on the unknown words in table 16, as this best captures the case where distributional evidence is more heavily relied upon.

Interestingly, we do not find the same results as with the frame-based classification. Here, the *noun form* mapping outperforms the *noun type* one. In other words, a mapping which we thought was more distributionally preferable—based on both internal and external criteria—results in worse accuracy. To see why this is, consider the fact that, in this tagging experiment, the most frequent confusable tags involving nouns are NN (common noun) and NNP (proper noun), which are conflated with the noun form, but not the noun type, mapping. Thus, mapping by noun form obtains a higher accuracy.

	Noun type	Noun form
Finiteness	96.27%	96.30%
Verb form	96.65%	**96.68%**

Table 15: Tagging accuracy

	Noun type	Noun form
Finiteness	73.18%	73.99%
Verb form	72.36%	73.16%

Table 16: Tagging accuracy for unknown words

This hearkens back to the point made earlier: a tagger is tagging all contexts, and many NN/NNP words appear only vary rarely, in contexts which do not recur. This shows, first of all, the limitation of using a POS tagger in order to test the distributional quality of a tagset. The tagger guesses in each context which tag is correct, but it is not explicitly marking a context as fitting a particular distributional frame. Secondly, it appears that at least in some tests, tagging accuracy is easier for some distinctions, but these are precisely the distinctions which it needs to be making. Once again, both internal and external criteria are needed when evaluating a tagset and a tagset mapping.

7. Summary and Outlook

In this paper, we have considered a range of corpora and tagsets, examining different tagset mappings and their distributional properties. There are various takeaway points: 1) Using frequent frames, or similar purely distributional tests, allows one to test how distributional a tagset is, in a way which is more reliable than a POS tagger. 2) When evaluating POS tagging or category induction methods involving a mapping to a simpler tagset, one should report a measurement of external quality; we specifically recommend one which records the number of ambiguities lost in the lexicon. This is especially important considering the differences in tagging accuracy for different mappings (cf. section 6.2.). 3) Tagset mappings can be done in a systematic "top-down" way, albeit limited by the original categories in the tagset, but automatic mappings show that better mappings are indeed possible.

Future work can go in a number of directions. Clearly, one can examine other corpora and tagsets, especially for morphologically-richer languages with more distinctions. In that process, there is much room for improvement in automatically defining mappings which can be used for a range of experiments. On that note, one should consider revisiting category induction experiments, to see what the quality of the methods are when considering mappings that do not lose as many linguistically-important distinctions.

References

Bosco, Cristina, Vincenzo Lombardo, Daniela Vassallo and Leonardo Lesmo (2000). Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of LREC 2000*. Athens, Greece, pp. 99–105.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). The TIGER Treebank. In *Proceedings of TLT-02*. Sozopol, Bulgaria.

Brants, Thorsten (1997). Internal and External Tagsets in Part-of-Speech Tagging. In *Proceedings of Eurospeech*. Rhodes, Greece.

Brants, Thorsten (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*. Seattle, WA.

Chanev, Atanas (2005). Portability of Dependency Parsing Algorithms - An Application for Italian. In *Proceedings of TLT-05*. Barcelona, Spain.

Chemla, E., T. H. Mintz, S. Bernal and A. Christophe (2009). Categorizing words using 'Frequent Frames': What cross-linguisic analyses reveal about core principles. *Developmental Science* 12(3), 396–406.

Clark, Alexander (2003). Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of EACL-03*. Budapest.

Déjean, Hervé (2000). How to Evaluate and Compare Tagsets? A Proposal. In *Proceedings LREC-00*. Athens.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera and Holger Wunsch (2010, to appear). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* .

Dickinson, Markus and Charles Jochim (2008). A Simple Method for Tagset Comparison. In *Proceedings of LREC 2008*. Marrakech, Morocco.

Dickinson, Markus and Charles Jochim (2009). Categorizing Local Contexts as a Step in Grammatical Category Induction. In *Proceedings of the EACL'09 Workshop on Cognitive Aspects of Computational Language Acquisition*. Athens, Greece.

Dickinson, Markus and Marwa Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of TLT-8*. Milan, Italy.

Elworthy, David (1995). Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*. Dublin.

Goldberg, Yoav, Meni Adler and Michael Elhadad (2008). EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start). In *Proceedings of ACL-08*. Columbus, OH.

Goldwater, Sharon and Tom Griffiths (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL-07*. Prague.

Harris, Zellig (1954). Distributional Structure. *Word* 10(23), 146–162.

MacWhinney, Brian (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third edn.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*. CUP.

Marcus, M., Beatrice Santorini and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

Mintz, Toben H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* 30, 678–686.

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.

Mintz, Toben H. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek and R. M. Golinkoff (eds.), *Action Meets Word: How Children Learn Verbs*, New York: Oxford University Press, pp. 31–63.

Pîrvan, Felix and Dan Tufiş (2006). Tagsets Mapping and Statistical Training Data Cleaning-up. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy: ELRA - European Language Ressources Association, pp. 385–390.

Redington, Martin, Nick Chater and Steven Finch (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22(4), 425–469.

Sampson, Geoffrey (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Smith, Noah A. and Jason Eisner (2005). Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of ACL'05*. Ann Arbor, MI.

Toutanova, Kristina and Mark Johnson (2008). A Bayesian LDA-based Model for Semi-Supervised Part-of-speech Tagging. In *Proceedings of NIPS 2008*. Vancouver.

Xiao, L., X. Cai and T. Lee (2006). The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. In *The Seventh Tokyo Conference on Psycholinguistics*. Keio University.

Zeman, Daniel (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC 2008*. Marrakech, Morocco.