A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing

John Lee
Department of Chinese,
Translation and Linguistics
City University of Hong Kong
jsylee@cityu.edu.hk

Jason Naradowsky, David A. Smith
Department of Computer Science
University of Massachusetts, Amherst
{narad,dasmith}@cs.umass.edu

Abstract

Most previous studies of morphological disambiguation and dependency parsing have been pursued independently. Morphological taggers operate on n-grams and do not take into account syntactic relations; parsers use the "pipeline" approach, assuming that morphological information has been separately obtained.

However, in morphologically-rich languages, there is often considerable interaction between morphology and syntax, such that neither can be disambiguated without the other. In this paper, we propose a discriminative model that jointly infers morphological properties and syntactic structures. In evaluations on various highly-inflected languages, this joint model outperforms both a baseline tagger in morphological disambiguation, and a pipeline parser in head selection.

1 Introduction

To date, studies of morphological analysis and dependency parsing have been pursued more or less independently. Morphological taggers disambiguate morphological attributes such as part-of-speech (POS) or case, without taking syntax into account (Hakkani-Tür et al., 2000; Hajič et al., 2001); dependency parsers commonly assume the "pipeline" approach, relying on morphological information as part of the input (Buchholz and Marsi, 2006; Nivre et al., 2007). This approach serves many languages well, especially those with less morphological ambiguity. In English, for example, accuracy of POS tagging has risen above 97% (Toutanova et al., 2003), and that of dependency parsing has reached the low nineties (Nivre et al., 2007). For these languages, there may be little to be gained to justify the computational cost of incorporating syntactic inference during the morphological tagging task; conversely, it is doubtful that errorful morphological information is a main cause of errors in English dependency parsing.

However, the pipeline approach seems more problematic for morphologically-rich languages with substantial interactions between morphology and syntax (Tsarfaty, 2006). Consider the Latin sentence, *Una dies omnis potuit praecurrere amantis*, 'One day was able to make up for all the lovers'[1]. As shown in Table 1, the adjective *omnis* ('all') is ambiguous in number, gender, and case; there are seven valid analyses. From the perspective of a finite-state morphological tagger, the most attractive analysis is arguably the singular nominative, since *omnis* is immediately followed by the singular verb *potuit* ('could'). Indeed, the baseline tagger used in this study did make this decision. Given its nominative case, the pipeline parser assigned the verb *potuit* to be its head; the two words form the typical subject-verb relation, agreeing in number.

Unfortunately, as shown in Figure 1, the word *omnis* in fact modifies the noun *amantis*, at the end of the sentence. As a result, despite the distance between them, they must agree in number, gender and case, i.e., both must be plural masculine (or feminine) accusative. The pipeline parser, acting on the input that *omnis* is nominative, naturally did not see

[1] Taken from poem 1.13 by Sextus Propertius, English translation by Katz (2004).

Latin	*Una*		*dies*		*omnis*			*potuit*	*praecurrere*	*amantis*	
English	one		day		all			could	to surpass	lovers	
Number	**sg**	pl	**sg**	pl	sg	sg	**pl**	**sg**	**-**	sg	**pl**
Gender	**f**	n	m/**f**	m/f	m/f	m/f/n	**m**/f	**-**	**-**	m/f/n	**m**/f
Case	**nom**/ab	nom/acc	**nom**	nom/acc	nom	gen	**acc**	**-**	**-**	gen	**acc**

Table 1: The Latin sentence "*Una dies omnis potuit praecurrere amantis*", meaning 'One day was able to make up for all the lovers', shown with glosses and possible morphological analyses. The correct analyses are shown in bold. The word *omnis* has 7 possible combinations of number, gender and case, while *amantis* has 5. Disambiguation partly depends on establishing *amantis* as the head of *omnis*, and so the two must agree in all three attributes.

this agreement, and therefore did not consider this syntactic relation likely.

Such a dilemma is not uncommon in languages with relatively free word order. On the one hand, it appears difficult to improve morphological tagging accuracy on words like *omnis* without syntactic knowledge; on the other hand, a parser cannot reliably disambiguate syntax unless it has accurate morphological information, in this example the agreement in number, gender, and case.

In this paper we propose to attack this chicken-and-egg problem with a discriminative model that jointly infers morphological and syntactic properties of a sentence, given its words as input. In evaluations on various highly-inflected languages, the model outperforms both a baseline tagger in morphological disambiguation, and a pipeline parser in head selection.

After a description of previous work (§2), the joint model (§3) will be contrasted with the baseline pipeline model (§4). Experimental results (§5-6) will then be presented, followed by conclusions and future directions.

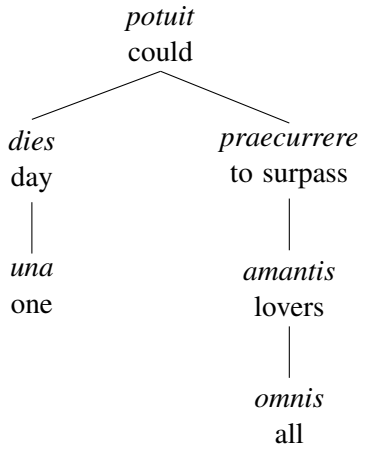

Figure 1: Dependency tree for the sentence "*Una dies omnis potuit praecurrere amantis*". The word *omnis* is an adjective modifying the noun *amantis*. This information is key to the morphological disambiguation of both words, as shown in Table 1.

2 Previous Work

Since space does not allow a full review of the vast literature on morphological analysis and parsing, we focus only on past research involving joint morphological and syntactic inference (§2.1); we then discuss Latin (§2.2), a language representative of the challenges that motivated our approach.

2.1 Joint Morphological and Syntactic Inference

Most previous work in morphological disambiguation, even when applied on morphologically complex languages with relatively free word order, such as Turkish (Hakkani-Tür et al., 2000) and Czech (Hajič et al., 2001), did not consider syntactic relationships between words. In the literature on data-driven parsing, two recent studies attempted joint inference on morphology and syntax, and both considered phrase-structure trees for Modern Hebrew (Cohen and Smith, 2007; Goldberg and Tsarfaty, 2008).

The primary focus of morphological processing in Modern Hebrew is splitting orthographic words into morphemes: clitics such as prepositions, pronouns, and the definite article must be separated from the core word. Each of the resulting morphemes is then tagged with an atomic "part-of-speech" to indicate word class and some morphological features. Similarly, the English POS tags in the Penn Treebank combine word class information with morphologi-

cal attributes such as "plural" or "past tense".

Cohen and Smith (2007) separately train a discriminative conditional random field (CRF) for segmentation and tagging, and a generative probabilistic context-free grammar (PCFG) for parsing. At decoding time, the two models are combined as a product of experts. Goldberg and Tsarfaty (2008) propose a generative joint model. This paper is the first to use a fully discriminative model for joint morphological and syntactic inference on dependency trees.

2.2 Latin

Unlike Modern Hebrew, Latin does not require extensive morpheme segmentation[2]. However, it does have a relatively free word order, and is also highly inflected, with each word having up to nine morphological attributes, listed in Table 2. In addition to its absolute numbers of cases, moods, and tenses, Latin morphology is *fusional*. For instance, the suffix $-is$ in *omnis* cannot be segmented into morphemes that separately indicate gender, number, and case. According to the Latin morphological database encoded in MORPHEUS (Crane, 1991), 30% of Latin nouns can be parsed as another part-of-speech, and on average each has 3.8 possible morphological interpretations.

We know of only one previous attempt in data-driven dependency parsing for Latin (Bamman and Crane, 2008), with the goal of constructing a dynamic lexicon for a digital library. Parsing is performed using the usual pipeline approach, first with the TreeTagger analyzer (Schmid, 1994) and then with a state-of-the-art dependency parser (McDonald et al., 2005). Head selection accuracy was 61.49%, and rose to 64.99% with oracle morphological tags. Of the nine morphological attributes, gender and especially case had the lowest accuracy. This observation echoes the findings for Czech (Smith et al., 2005), where case was also the most difficult to disambiguate.

Attribute	Values
Part-of-speech (POS)	noun, verb, participle, adjective, adverb, conjunction, preposition, pronoun, numeral, interjection, exclamation, punctuation
Person	first, second, third
Number	singular, plural
Tense	present, imperfect, perfect, pluperfect, future perfect, future
Mood	indicative, subjunctive, infinitive, imperative, participle, gerund, gerundive, supine
Voice	active, passive
Gender	masculine, feminine, neuter
Case	nominative, genitive, dative, accusative, ablative, vocative, locative
Degree	comparative, superlative

Table 2: Morphological attributes and values for Latin. Ancient Greek has the same attributes; Czech and Hungarian lack some of them. In all categories except POS, a value of *null* ('-') may also be assigned. For example, a noun has '-' for the tense attribute.

3 Joint Model

This section describes a model that jointly infers morphological and syntactic properties of a sentence. It will be presented as a graphical model, starting with the variables and then the factors, which represents constraints on the variables. Let n be the number of words and m be the number of possible values for a morphological attribute. The variables are:

- WORD: the n words w_1,...,w_n of the input sentence, all observed.
- TAG: $O(nm)$ boolean variables[3] $T_{a,i,v}$, corresponding to each value of the morphological attributes listed in Table 2. $T_{a,i,v} = true$ when the word w_i has value v as its morphological attribute a. In Figure 2, CASE$_{3,acc}$ is the shorthand representing the variable $T_{case,3,acc}$. It is set to $true$ since the word w_3 has the accusative case.
- LINK: $O(n^2)$ boolean variables $L_{i,j}$ corresponding to a possible link between each pair

[2]Except for enclitics such as *-que*, *-ve*, and *-ne*, but their segmentation is rather straightforward compared to Modern Hebrew or other Semitic languages.

[3]The TAG variables were actually implemented as multinomials, but are presented here as booleans for ease of understanding.

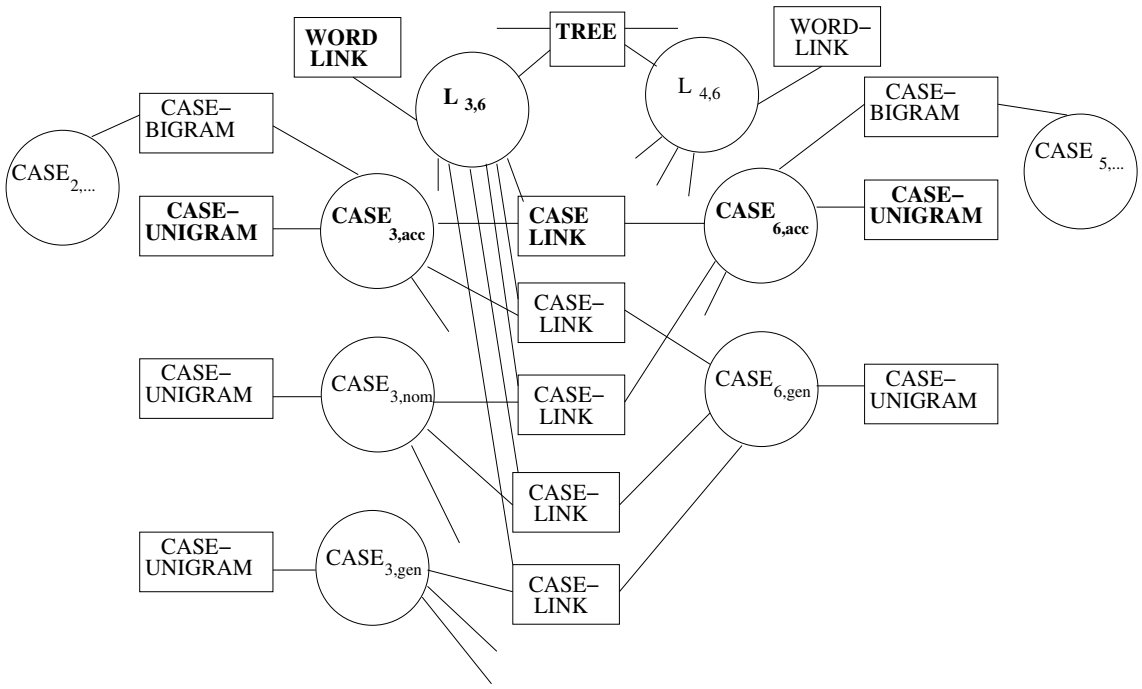

Figure 2: The joint model (§3) depicted as a graphical model. The variables, all boolean, are represented by circles and are bolded if their correct values are $true$. Factors are represented by rectangles and are bolded if they fire. For clarity, this graph shows only those variables and factors associated with one pair of words (i.e., w_3=*omnis* and w_6=*amantis*) and with one morphological attribute (i.e., case). The variables $L_{3,6}$, CASE$_{3,acc}$ and CASE$_{6,acc}$ are bolded, indicating that w_3 and w_6 are linked and both have the accusative case. The ternary factor CASE-LINK, that connects to these three variable, therefore fires.

of words[4]. $L_{i,j} = true$ when there is a dependency link from the word w_i to the word w_j. In Figure 2, the variable $L_{3,6}$ is set to $true$ since there is a dependency link between the words w_3 and w_6.

We define a probability distribution over all joint assignments $\mathcal{A}$ to the above variables,

$$p(\mathcal{A}) = \frac{1}{Z} \prod_k F_k(\mathcal{A}) \quad (1)$$

where Z is a normalizing constant. The assignment $\mathcal{A}$ is subject to a hard constraint, represented in Figure 2 as TREE, requiring that the values of the LINK variables must yield a tree, which may be non-projective. The factors $F_k(\mathcal{A})$ represent soft constraints evaluating various aspects of the "goodness" of the tree structure implied by $\mathcal{A}$. We say a factor "fires" when all its neighboring variables are $true$ and it evaluates to a non-negative real number; otherwise, it evaluates to 1 and has no effect on the product in equation (1). Soft constraints in the model are divided into **local** and **link** factors, to which we now turn.

[4]Variables for link labels can be integrated in a straightforward manner, if desired.

3.1 Local Factors

The local factors consult either one word or two neighboring words, and their morphological attributes. These factors express the desirability of the assignments of morphological attributes based on local context. There are three types:

- TAG-UNIGRAM: There are $O(nm)$ such unary factors, each instance of which is connected to a TAG variable. The factor fires when $T_{a,i,v}$ is $true$. The features consist of the value v of the morphological attribute concerned, combined with the word identity of w_i, with backoff using all suffixes of the word. The CASE-UNIGRAM factors shown in Figure 2 are examples of this family of factors.

- TAG-BIGRAM: There are $O(nm^2)$ of such binary factors, each connected to the TAG variables of a pair of neighboring words. The factor fires when T_{a,i,v_1} and $T_{a,i+1,v_2}$ are both $true$. The CASE-BIGRAM factors shown in Figure 2 are examples of this family of factors.

- TAG-CONSISTENCY: For each word, the TAG variables representing the possible POS values are connected to those representing the values of other morphological attributes, yielding $O(nm^2)$ binary factors. They fire when T_{pos,i,v_1} and T_{a,i,v_2} are both $true$. These factors are intended to discourage inconsistent assignments, such as a non-null tense for a noun.

It is clear that so far, none of these factors are aware of the morphological agreement between *omnis* and *amantis*, crucial for inferring their syntactic relation. We now turn our attention to link factors, which serve this purpose.

3.2 Link Factors

The link factors consult all pairs of words, possibly separated by a long distance, that may have a dependency link. These factors model the likelihood of such a link based on the word identities and their morphological attributes:

- WORD-LINK: There are $O(n^2)$ such unary factors, each connected to a LINK variable, as shown in Figure 2. The factor fires when $L_{i,j}$ is $true$. Features include various combinations of the word identities of the parent w_i and child w_j, and 5-letter prefixes of these words, replicating the so-called "basic features" used by McDonald et al. (2005).

- POS-LINK: There are $O(n^2m^2)$ such ternary factors, each connected to the variables $L_{i,j}$, T_{i,pos,v_i} and T_{j,pos,v_j}. It fires when all three are $true$ or, in other words, when the parent word w_i has POS v_i, and the child w_j has POS v_j. Features replicate all the so-called "basic features" used by McDonald et al. (2005) that involve POS. These factors are not shown in Figure 2, but would have exactly the same structure as the CASE-LINK factors.

  Beyond these basic features, McDonald et al. (2005) also utilize POS trigrams and POS 4-grams. Both include the POS of two linked words, w_i and w_j. The third component in the trigrams is the POS of each word w_k located between w_i and w_j, $i < k < j$. The two additional components that make up the 4-grams are subsets of the POS of words located to the immediate left and right of w_i and w_j.

  If fully implemented in our joint model, these features would necessitate two separate families of link factors: $O(n^3m^3)$ factors for the POS trigrams, and $O(n^2m^4)$ factors for the POS 4-grams. To avoid this substantial increase in model complexity, these features are instead approximated: the POS of all words involved in the trigrams and 4-grams, except those of w_i and w_j, are regarded as fixed, their values being taken from the output of a morphological tagger (§4.1), rather than connected to the appropriate TAG variables. This approximation allows these features to be incorporated in the POS-LINK factors.

- MORPH-LINK: There are $O(n^2m^2)$ such ternary factors, each connected to the variables $L_{i,j}$, T_{i,a,v_i} and T_{j,a,v_j}, for every attribute a other than POS. The factor fires when all three variables are $true$, and both v_i and v_j are non-null; i.e., it fires when the parent word w_i has v_i as its morphological attribute a, and the child w_j has v_j. Features include the combination of v_i and v_j themselves, and agreement between them. The CASE-LINK factors in Figure 2 are an example of this family of factors.

4 Baselines

To ensure a meaningful comparison with the joint model, our two baselines are both implemented in the same graphical model framework, and trained with the same machine-learning algorithm. Roughly speaking, they divide up the variables and factors of the joint model and train them separately. For morphological disambiguation, we use the baseline tagger described in §4.1. For dependency parsing, our baseline is a "pipeline" parser (§4.2) that infers syntax upon the output of the baseline tagger.

4.1 Baseline Morphological Tagger

The tagger is a graphical model with the WORD and TAG variables, connected by the local factors TAG-UNIGRAM, TAG-BIGRAM, and TAG-CONSISTENCY, all used in the joint model (§3).

4.2 Baseline Dependency Parser

The parser has no local factors, but has the same variables as the joint model and the same features from all three families of link factors (§3). However, since it takes as input the morphological attributes predicted by the tagger, the TAG variables are now observed. This leads to a change in the structure of the link factors — all features from the POS-LINK factors now belong to the WORD-LINK factors, since the POS of all words are observed. In short, the features of the parser are a replication of (McDonald et al., 2005), but also extended beyond POS to the other morphological attributes, with the features in the MORPH-LINK factors incorporated into WORD-LINK for similar reasons.

5 Experimental Set-up

5.1 Data

Our evaluation focused on the Latin Dependency Treebank (Bamman and Crane, 2006), created at the Perseus Digital Library by tailoring the Prague Dependency Treebank guidelines for the Latin language. It consists of excerpts from works by eight Latin authors. We randomly divided the 53K-word treebank into 10 folds of roughly equal sizes, with an average of 5314 words (347 sentences) per fold. We used one fold as the development set and performed cross-validation on the other nine.

To measure how well our model generalizes to other highly-inflected, relatively free-word-order languages, we considered Ancient Greek, Hungarian, and Czech. Their respective datasets consist of 8000 sentences from the Ancient Greek Dependency Treebank (Bamman et al., 2009), 5800 from the Hungarian Szeged Dependency Treebank (Vincze et al., 2010), and a subset of 3100 from the Prague Dependency Treebank (Böhmová et al., 2003).

5.2 Training

We define each factor in (1) as a log-linear function:

$$F_k(\mathcal{A}) = \exp \sum_h \theta_h f_h(\mathcal{A}, W, k) \qquad (2)$$

Given an assignment $\mathcal{A}$ and words W, f_h is an indicator function describing the presence or absence of the feature, and θ_h is the corresponding set of weights learned using stochastic gradient ascent, with the gradients inferred by loopy belief propagation (Smith and Eisner, 2008). The variance of the Gaussian prior is set to 1. The other two parameters in the training process, the number of belief propagation iterations and the number of training rounds, were tuned on the development set.

5.3 Decoding

The output of the joint model is the assignment to the TAG and LINK variables. Loopy belief propagation (BP) was used to calculate the posterior probabilities of these variables. For TAG, we emit the tag with the highest posterior probability as computed by sum-product BP. We produced head attachments by first calculating the posteriors of the LINK variables with BP and then passing them to an edge-factored tree decoder. This is equivalent to minimum Bayes risk decoding (Goodman, 1996), which is used by Cohen and Smith (2007) and Smith and Eisner (2008). This MBR decoding procedure enforces the hard constraint that the output be a tree but sums over possible morphological assignments.[5]

5.4 Reducing Model Complexity

In principle, the joint model should consider every possible combination of morphological attributes for every word. In practice, to reduce the complexity of the model, we used a pre-existing morphological database, MORPHEUS (Crane, 1991), to constrain the range of possible values of the attributes listed in Table 2; more precisely, we add a hard constraint, requiring that assignments to the TAG variables be compatible with MORPHEUS. This constraint significantly reduces the value of m in the big-O notation

[5] This approach to nuisance variables has also been used effectively for parsing with tree-substitution grammars, where several derived trees may correspond to each derivation tree, and parsing with PCFGs with latent annotations.

Model Attr. ↓	Tagger all	Joint all	Tagger non-null	Joint non-null
POS	94.4	94.5	94.4	94.5
Person	99.4	99.5	97.1	97.6
Number	95.3	95.9	93.7	94.5
Tense	98.0	98.2	93.2	93.9
Mood	98.1	98.3	93.8	94.4
Voice	98.5	98.6	95.3	95.7
Gender	93.1	93.9	87.7	89.1
Case	89.3	90.0	79.9	81.2
Degree	99.9	99.9	86.4	90.8
UAS	61.0	61.9	—	—

Table 3: **Latin** morphological disambiguation and parsing. For some attributes, such as degree, a substantial portion of words have the *null* value. The non-null columns provides a sharper picture by excluding these "easy" cases. Note that POS is never *null*.

for the number of variables and factors described in §3. To illustrate the effect, the graphical model of the sentence in Table 1, whose six words are all covered by the database, has 1,866 factors; without the benefit of the database, the full model would have 31,901 factors.

The MORPHEUS database was automatically generated from a list of stems, inflections, irregular forms and morphological rules. It covers about 99% of the distinct words in the Latin Dependency Treebank. At decoding time, for each fold, the database is further augmented with tags seen in training data. After this augmentation, an average of 44 words are "unseen" in each fold.

Similarly, we constructed morphological dictionaries for Czech, Ancient Greek, and Hungarian from words that occurred at least five times in the training data; words that occurred fewer times were unrestricted in the morphological attributes they could take on.

6 Experimental Results

We compare the performance of the pipeline model (§4) and the joint model (§3) on morphological disambiguation and unlabeled dependency parsing.

Model Attr. ↓	Tagger all	Joint all	Tagger non-null	Joint non-null
POS	95.5	95.7	95.5	95.7
Person	98.4	98.8	93.5	95.6
Number	91.2	92.3	87.0	88.4
Tense	98.4	98.8	92.7	96.1
Voice	98.5	98.7	93.2	95.8
Gender	86.6	87.9	75.6	78.0
Case	84.1	85.6	74.3	76.5
Degree	97.9	98.0	90.1	90.1
UAS	67.4	68.7	—	—

Table 4: **Czech** morphological disambiguation and parsing. As with Latin, the model is least accurate with noun/adjective categories of gender number, and case, particularly when considering only words whose true value is non-null for those attributes. Joint inference with syntactic features improves accuracy across the board.

Model Attr. ↓	Tagger all	Joint all	Tagger non-null	Joint non-null
POS	94.9	95.7	94.9	95.7
Person	98.7	99.0	92.2	94.6
Number	97.4	97.9	96.5	97.1
Tense	96.8	97.2	84.1	86.8
Mood	97.9	98.3	91.4	93.2
Voice	97.8	98.0	91.3	92.4
Gender	95.4	96.1	90.7	91.9
Case	95.9	96.3	92.0	92.6
Degree	99.8	99.9	33.3	55.6
UAS	68.0	70.5	—	—

Table 5: **Ancient Greek** morphological disambiguation and parsing. Noun/adjective morphology is more accurate, but verbal morphology is more problematic.

Model Attr. ↓	Tagger all	Joint all	Tagger non-null	Joint non-null
POS	95.8	95.8	95.8	95.8
Person	98.5	98.6	94.9	94.1
Number	97.4	97.5	96.8	96.6
Tense	98.9	99.3	97.2	97.3
Mood	98.7	99.2	95.8	97.3
Case	96.7	97.0	94.5	94.9
Degree	97.9	98.1	87.5	88.6
UAS	78.2	78.8	—	—

Table 6: **Hungarian** morphological disambiguation and parsing. The agglutinative morphological system makes local cues more effective, but syntactic information helps in almost all categories.

6.1 Morphological Disambiguation

As seen in Table 3, the joint model outperforms[6] the baseline tagger in all attributes in Latin morphological disambiguation. Among words not covered by the morphological database, accuracy in POS is slightly better, but lower for case, gender and number.

The joint model made the most gains on adjectives and participles. Both parts-of-speech are particularly ambiguous: according to MORPHEUS, 43% of the adjectives can be interpreted as another POS, most frequently nouns; while participles have an average of 5.5 morphological interpretations. Both also often have identical forms for different genders, numbers and cases. In these situations, syntactic considerations help nudge the joint model to the correct interpretations.

Experiments on the other three languages bear out similar results: the joint model improves morphological disambiguation. The performance of Czech (Table 4) exhibits the closest analogue to Latin: gender, number, and case are much less accurately predicted than are the other morphological attributes. Like Latin, Czech lacks definite and indefinite articles to provide high-confidence cues for noun phrase boundaries.

The Ancient Greek treebank comprises both archaic texts, before the development of a definite article, and later classic Greek, which has a definite article; Hungarian has both a definite and an indefinite article. In both languages (Tables 5 and 6), noun and adjective gender, number, and case are more accurately predicted than in Czech and Latin. The verbal system of ancient Greek, in contrast, is more complex than that of the other languages, so mood, voice, and tense accuracy are lower.

6.2 Dependency Parsing

In addition to morphological disambiguation, we also measured the performance of the joint model on dependency parsing of Latin and the other languages. The baseline pipeline parser (§4.2) yielded 61.00% head selection accuracy (i.e., unlabeled attachment score, UAS), outperformed[7] by the joint model at 61.88%. The joint model showed similar improvements in Ancient Greek, Hungarian, and Czech.

Wrong decisions made by the baseline tagger often misled the pipeline parser. For adjectives, the example shown in Table 1 and Figure 1 is a typical scenario, where an accusative adjective was tagged as nominative, and was then misanalyzed by the parser as modifying a verb (as a subject) rather than modifying an accusative noun. For participles modifying a noun, the wrong noun was often chosen based on inaccurate morphological information. In these cases, the joint model, entertaining all morphological possibilities, was able to find the combination of links and morphological analyses that are collectively more likely.

The accuracy figures of our baselines are comparable, but not identical, to their counterparts reported in (Bamman and Crane, 2008). The differences may partially be attributed to the different morphological tagger used, and the different learning algorithm, namely Margin Infused Relaxed Algorithm (MIRA) in (McDonald et al., 2005) rather than maximum likelihood. More importantly, the Latin Dependency Treebank has grown from about 30K at the time of the previous work to 53K at present, resulting in significantly different training and testing material.

Gold Pipeline Parser When given perfect morphological information, the Latin parser performs at 65.28% accuracy in head selection. Despite the oracle morphology, the head selection accuracy is still below other languages. This is hardly surprising, given the relatively small training set, and that the "the most difficult languages are those that combine a relatively free word order with a high degree of inflection", as observed at the recent dependency parsing shared task (Nivre et al., 2007); both of these are characteristics of Latin.

A particularly troublesome structure is coordination; the most frequent link errors all involve either a parent or a child as a conjunction. In a list of words, all words and coordinators depend on the final coordinator. Since the factors in our model consult only one link at a time, they do not sufficiently capture this kind of structures. Higher-order features, particularly those concerned with links with grandparents and siblings, have been shown to benefit dependency

[6] The differences are statistically significant in all ($p < 0.01$ by McNemar's Test) but POS ($p = 0.5$).

[7] Significant at $p < e^{-11}$ by McNemar's Test.

parsing (Smith and Eisner, 2008) and may be able to address this issue.

7 Conclusions and Future Work

We have proposed a discriminative model that jointly infers morphological properties and syntactic structures. In evaluations on various highly-inflected languages, this joint model outperforms both a baseline tagger in morphological disambiguation, and a pipeline parser in head selection.

This model may be refined by incorporating richer features and improved decoding. In particular, we would like to experiment with higher-order features (§6), and with maximum *a posteriori* decoding, via max-product BP or (relaxed) integer linear programming. Further evaluation on other morphological systems would also be desirable.

Acknowledgments

We thank David Bamman and Gregory Crane for their feedback and support. Part of this research was performed by the first author while visiting Perseus Digital Library at Tufts University, under the grants *A Reading Environment for Arabic and Islamic Culture*, Department of Education (P017A060068-08) and *The Dynamic Lexicon: Cyberinfrastructure and the Automatic Analysis of Historical Languages*, National Endowment for the Humanities (PR-50013-08). The latter two authors were supported by Army prime contract #W911NF-07-1-0216 and University of Pennsylvania subaward #103-548106; by SRI International subcontract #27-001338 and ARFL prime contract #FA8750-09-C-0181; and by the Center for Intelligent Information Retrieval. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

David Bamman and Gregory Crane. 2006. The Design and Use of a Latin Dependency Treebank. *Proc. Workshop on Treebanks and Linguistic Theories (TLT)*. Prague, Czech Republic.

David Bamman and Gregory Crane. 2008. Building a Dynamic Lexicon from a Digital Library. *Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*. Pittsburgh, PA.

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. *Proc. Workshop on Treebanks and Linguistic Theories (TLT)*.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The PDT: a 3-level Annotation Scenario. In *Treebanks: Building and Using Parsed Corpora*, A. Abeillé (ed). Kluwer.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. *Proc. CoNLL*. New York, NY.

Shay B. Cohen and Noah A. Smith. 2007. Joint Morphological and Syntactic Disambiguation. *Proc. EMNLP-CoNLL*. Prague, Czech Republic.

Gregory Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing* 6(4):243–245.

Yoav Goldberg and Reut Tsarfaty. 2008. A Single Generative Model for Joint Morphological Segmentation and Syntactic Parsing. *Proc. ACL*. Columbus, OH.

Joshua Goodman. 1996. Parsing Algorithms and Metrics. *Proc. ACL*.

J. Hajič, P. Krbec, P. Květoň, K. Oliva, and V. Petkevič. 2001. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. *Proc. ACL*.

D. Z. Hakkani-Tür, K. Oflazer, and G. Tür. 2000. Statistical Morphological Disambiguation for Agglutinative Languages. *Proc. COLING*.

Vincent Katz. 2004. *The Complete Elegies of Sextus Propertius*. Princeton University Press, Princeton, NJ.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jana Hajič. 2005. Non-projective Dependency Parsing using Spanning Tree Algorithms. *Proc. HLT/EMNLP*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online Large-Margin Training of Dependency Parsers. *Proc. ACL*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. *Proc. CoNLL Shared Task Session of EMNLP-CoNLL*. Prague, Czech Republic.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. *Proc. International Conference on New Methods in Language Processing*. Manchester, UK.

Noah A. Smith, David A. Smith and Roy W. Tromble. 2005. Context-Based Morphological Disambiguation with Random Fields. *Proc. HLT/EMNLP*. Vancouver, Canada.

David Smith and Jason Eisner. 2008. Dependency Parsing by Belief Propagation. *Proc. EMNLP*. Honolulu, Hawaii.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proc. HLT-NAACL*. Edmonton, Canada.

Reut Tsarfaty. 2006. Integrated Morphological and Syntactic Disambiguation for Modern Hebrew. *Proc. COLING-ACL Student Research Workshop*.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. *Proc. LREC*.