Learning Morphology with Morfette

Grzegorz Chrupała, Georgiana Dinu, Josef van Genabith

Dublin City University
Dublin 9, Ireland
gchrupala@computing.dcu.ie

Universität des Saarlandes
D-66041 Saarbrücken, Germany
dinu@coli.uni-sb.de

Dublin City University
Dublin 9, Ireland
josef@computing.dcu.ie

Abstract

Morfette is a modular, data-driven, probabilistic system which learns to perform joint morphological tagging and lemmatization from morphologically annotated corpora. The system is composed of two learning modules which are trained to predict morphological tags and lemmas using the Maximum Entropy classifier. The third module dynamically combines the predictions of the Maximum-Entropy models and outputs a probability distribution over tag-lemma pair sequences. The lemmatization module exploits the idea of recasting lemmatization as a classification task by using class labels which encode mappings from wordforms to lemmas. Experimental evaluation results and error analysis on three morphologically rich languages show that the system achieves high accuracy with no language-specific feature engineering or additional resources.

1. Introduction

This paper describes and evaluates the **Morfette** system for data-driven morphological analysis. Morphological analysis usually involves two subtasks: the assignment of morphological features to the wordform, and lemmatization.

Many data-driven approaches to morphology involve encoding morphological features as tags (henceforth *morpho-tags*), and use some sequence labeling method to assign *morpho-tag* sequences to sentences. In the case of morphologically rich inflectional or agglutinative languages, the classification decision is often constrained by the use of a morphological lexicon, or a finite-state morphological analyzer: in such systems the data-driven component is limited to performing morphological disambiguation rather than morphological analysis itself (Hajič and Hladká, 1998; Hajič, 2000; Tufiş, 1999; Tufiş and Dragomirescu, 2004; Ceauşu, 2006; Han and Palmer, 2004; Habash and Rambow, 2005; Hakkani-Tür et al., 2002; Yuret and Türe, 2006).

In an morphological disambiguation setting, lemmatization is simple: either the lexicon or the morphological analyzer already returns the correct lemma corresponding to each of the candidate analyses. The problematic cases are unknown words: most systems are able to guess the morpho-tag of an unknown word, but not the corresponding lemma. (Erjavec and Džeroski, 2004) solve the problem of lemmatizing unknown words by using a two stage architecture, first sentences are assigned morpho-tag sequences by a POS-tagger, and then an Inductive Logic Programming system assigns lemmas to unknown wordform-tag pairs.

(Chrupała, 2006) takes a different approach to lemmatization. His method automatically induces lemma-classes: they correspond to the *shortest edit script* between reversed wordforms and the corresponding lemmas. Then a standard classifier is used to "tag" words with their lemma-classes, from which the words' lemmas can be obtained by "executing" the edit script on the wordforms. Thus in this approach lemmatization becomes quite similar to POS tagging or morphological tagging.

In the current study we present a modular, data-driven model which performs both morphological tagging and lemmatization, i.e. it maps a sequence of wordforms of length n to the sequence of morpho-tag - lemma pairs:

$$M : \mathcal{W}^n \rightarrow (\mathcal{M} \times \Lambda)^n \quad (1)$$

We use a generic, language-independent feature-set in our models and investigate how well such an approach generalizes to three morphologically rich languages.

In Section 2. we present the architecture of our model, the features used and the search algorithm. In Section 3. we present experimental evaluation results for three languages and corpora. Section 4. contains the error analysis and finally Section 5. presents our conclusions and ideas for further improvements in data-driven morphological analysis.

2. The Morfette System

2.1. Architecture

The **Morfette** system is composed of two learning modules, one for morphological tagging and one for lemmatization, and one decoding module which searches for the best sequence of pairs of morphological tags and lemmas for an input sequence of wordforms. Both modules learn Maximum Entropy classifiers such as that described for POS tagging in (Ratnaparkhi, 1996). For the lemmatization model we use (Chrupała, 2006)'s method of inducing lemma classes.[1]

In his method the class assigned to a wordform - lemma pair is the corresponding *shortest edit script* (henceforth SES) between the two reversed strings (Myers, 1986). A SES stands for the shortest sequence of instructions (insertions or deletions) which transforms a string w into a string w'. For example, considering the strings $w = pidieron$ and $w' = pedir$, the corresponding SES is $\{\langle$D, i, 2 $\rangle$, $\langle$I, e, 3 $\rangle$, $\langle$D, e, 5 $\rangle$, $\langle$D, o, 7 $\rangle$, $\langle$D, n, 8 $\rangle\}$, where a triple such as $\langle$D, i, 2 $\rangle$ stands for *delete character i at position 2*. Since one needs to abstract away from the length of words as much as possible, and since inflection affects predominantly word endings, the strings are reversed prior to the computation of SESs. In this way, pairs such as $\langle pidieron, pedir\rangle$ and $\langle repitieron, repetir\rangle$ are correctly

[1] We do not, however, use the features or the SVM classifier as presented in that paper, as we found that such a configuration is impractically slow in practice and scales poorly.

assigned the same class, corresponding to their shared position in the verb inflection paradigm.

2.2. Features

In our architecture we can use arbitrary features of the focus word and the context sentence. We use a rather minimalistic and language independent feature set in the experiments in Section 3. This has the advantage of being very general and using very little domain expertise but obviously for maximum performance it is desirable to extend and refine it using language and domain specific features. Table 1 shows the features extracted for the morphological tagging and lemmatization models. Table 2 exemplifies the morpho-tagging feature extraction for the words in the example sentence in Figure 1.

For the morphological tagging model we extract the wordform, lemma, and morpho-tag of the preceding two words, suffixes of length 1-7 and prefixes of length 1-5 of the focus wordform, as well as the spelling pattern of the focus wordform (wordform features are lower-cased). The spelling pattern feature encodes character classes such as upper-case and lower-case letters, digits, hyphens, underscores and other punctuation. Additionally we use the wordform of one following token, the set of morpho-tags present in the training data for its wordform, and a concatenation of the part-of-speech component of the morpho-tags of the previous two words.

For the lemmatization model a similar but smaller feature set is used: wordform, morpho-tag, suffixes of length 1-7, prefixes of length 1-5 and spelling pattern of the focus word. The fact that we use the morpho-tag of the focus word as a feature for the lemmatizer is important for the search algorithm described in Section 2.3.

2.3. Search

Maximum entropy models trained on examples such the the ones shown above predict probability distributions over classes (i.e. morpho-tags or lemma-classes) for the current focus wordform given its context as encoded in the features. That is for a focus word w_i in context $c \in \mathcal{C}$ for each possible morpho-tag $m \in \mathcal{M}$ the morpho-tagging model gives $p(m|c)$, and for each possible lemma-class $l \in \mathcal{L}$ the lemmatization model gives $p(l|c, m)$. The context includes the focus wordform as well as the preceding and following wordforms in the same sentence.

The algorithm is a beam search which maintains a list of n-best sequences of $(m, l) \in \mathcal{M} \times \mathcal{L}$ (morpho-tag - lemma-class) pairs up to the current position in the input word sequence. The conditional probability of a candidate sequence for words $w_0..w_i$ is given by

$$P(m_0..m_i, l_0..l_i|w_0..w_i) = \quad (2)$$

$$= p(l_i|c_i, m_i)p(m_i|c_i)P(m_0..m_{i-1}, l_0..l_{i-1}|w_0..w_{i-1})$$

The search proceeds as follows: for focus word w_i there are n (n being the beam size) highest probability sequences $((m_0, l_0)..(m_{i-1}, l_{i-1}))$. For each of those sequences we obtain a morpho-tag probability distribution from the morpho-tagging model. For efficiency reasons we pre-prune this set of tags: given the list of tag probabilities $(m_0, p_0)..(m_j, p_j)$ sorted in decreasing order, we keep all the tags $m_0..m_i$ where p_i satisfies the condition:

$$p_i / \sum_{k=0}^{i} p_k < T,$$

where T is a threshold parameter. Each of the retained morpho-tags for word w_i is added to each candidate sequence and for each of those combinations we obtain lemma-class probability distribution from the lemmatization model. The lemma-class set is pruned according to the same method as for morpho-tags. The probability of candidate sequences is updated according to equation 2, and the n highest ranking candidate sequences for $w_0..w_i$ are retained as the algorithm proceeds to word w_{i+1}.

3. Evaluation

For evaluation we chose three morphologically rich languages for which we have expertise to perform error analysis. We have not tuned the features or parameters of our system to any particular dataset. At this stage we are not interested in necessarily improving on the best published results for a particular language; rather we want to see how well the system performs with a minimalistic feature set and no language-dependent engineering effort and identify the main source of mistakes for each language.

We use the following data sets:

- Romanian: MULTEXT-EAST corpus (Erjavec, 2004), approx. 13,500 tokens (chapters 1-3) as a test set, approx. 11,800 tokens (chapters 5 and 6) for development and 88,000 tokens (chapters 7-23) for training.
- Spanish: CESS-ECE treebank (Martí et al., 2007), approx. 10,000 tokens each for test and development set, and approx. 168,000 tokens for the FULL training set, and approx. 70,000 for the SMALL training set.
- Polish: Korpus Słownika Frekwencyjnego (IPI PAN)[2], 10,000 tokens each for test and development sets, and approx. 219,000 for FULL training set, and approx. 70,000 for the SMALL training set.

The SMALL training set was used in order to be able to have similar training set sizes across the three languages. Additionally for Polish and Spanish the FULL set contains all the available data. No larger-size training data is available in the Romanian corpus.

For all the experiments reported in the following sections a beam size of 3 was used, with the prepruning threshold set to 0.3: validation on the development sets showed that those settings give good results for all the languages.

Table 3 shows the evaluation results for the SMALL training set for all three languages. Table 4 shows the results for Spanish and Polish, for which there is a larger training set available. More data is clearly beneficial: the scores improve substantially for both languages.

Both the morphological tagging and lemmatization scores for Polish are lower than for the other two languages: this is to be expected for a Slavic language with a rich inflection

[2] Available at http://korpus.pl/index.php?page=download

Wordform	În	pereţii	**boxei**	erau	trei	orificii
Lemma	în	perete	**boxă**	fi	trei	orificiu
Morpho-tag	Spsa	Ncmpry	**Ncfsoy**	Vmii3p	Mc-p-l	Ncfp-n
Gloss	*In*	*the walls of the cubicle*		*there were*	*three*	*orifices*

Figure 1: Example of a sentence in the Romanian MULTEXT-EAST corpus

Feature notation	Description
Morpho-tagging model	
f_0	Lowercased wordform of the focus token
$s_n(f_0),\ n = 1 \cdots 7$	Suffixes of length n
$p_n(f_0),\ n = 1 \cdots 5$	Prefixes of length n
$sp(F_0)$	Spelling pattern of the (non-lowercased) wordform
$s_1(m_{-2}) \oplus s_1(m_{-1})$	Concatenation of the first element of the two previous morpho-tags
f_{-2}, f_{-1}, f_1	Lowercased wordform of two previous tokens and of one following token
m_{-2}, m_{-1}	(Predicted) Morpho-tag of two previous tokens
l_{-2}, l_{-1}	(Predicted) Lemma of two previous tokens
m_{train_1}	Set of morpho-tags seen in training data for wordform of next token
Lemmatization model	
f_0	Lowercased wordform of the focus token
$s_n(f_0),\ n = 1 \cdots 7$	Suffixes of length n
$p_n(f_0),\ n = 1 \cdots 5$	Prefixes of length n
m_0	(Predicted) Morpho tag
$sp(F_0)$	Spelling pattern of the (non-lowercased) wordform

Table 1: Features used for the morphological tagging and lemmatization models

All words			
	Morpho-tagging	Lemmatization	Joint
Romanian	96.83	97.78	96.08
Spanish	94.33	97.84	93.83
Polish	81.87	93.29	81.19
Unseen words			
	Morpho-tagging	Lemmatization	Joint
Romanian	86.68	82.88	78.50
Spanish	74.79	89.20	71.26
Polish	61.93	76.88	59.17

Table 3: Evaluation results with SMALL training sets

All words			
	Morpho-tagging	Lemmatization	Joint
Spanish	95.40 (+1.07)	98.52 (+0.68)	95.02 (+1.19)
Polish	84.91 (+3.04)	95.55 (+2.26)	84.44 (+3.25)
Unseen words			
	Morpho-tagging	Lemmatization	Joint
Spanish	75.71 (+4.22)	91.22 (+2.74)	71.84 (+3.99)
Polish	65.87 (+4.33)	81.11 (+4.49)	63.16 (+4.33)

Table 4: Evaluation results with a FULL training set. Numbers in brackets indicate accuracy improvement over the same model trained on the SMALL training set

and high ambiguity. These properties are reflected in Table 5, which shows the average number of morpho-tag classes per token for each of the three languages, as well as the percentage of tokens for which the lemma is identical to the wordform.

	Avg. morpho-tags	Id. lemmas
Romanian	1.16	58.72%
Spanish	1.46	66.73%
Polish	2.23	44.44%

Table 5: Average morpho-tag ambiguity per token and percentage of tokens with lemmas identical to wordforms.

For comparison we have experimented with simple alternative data-driven methods for morpho-tagging and lemmatization. The accuracy achieved using these methods can be considered a non-trivial baseline for our joint model. The BASELINE model we constructed is composed of the following two components working in a pipeline:

- Morphological tagging: A tagger is obtained from training material using the MBT memory-based tagger generator (Daelemans et al., 2007).
- Lemmatization: For each wordform in the test set the morpho-tag predicted by MBT is retrieved. If the word - morpho-tag pair has been encountered in the training set, then it is assigned its predominant lemma; otherwise a lemma identical to the wordform is assigned.

Table 6 shows the performance of these two methods on morphological tagging and lemmatization as well as their joint accuracy on morpho-tag - lemma prediction. The results reported are obtained using the FULL training sets. Numbers in parentheses compare the results with the accuracy scores obtained with our model.

f_0	$sp(F_0)$	$p_1(f_0)$	$p_2(f_0)$	$p_3(f_0)$	$p_4(f_0)$	$p_5(f_0)$	$s_1(f_0)$	$s_2(f_0)$	$s_3(f_0)$	$s_4(f_0)$	$s_5(f_0)$	$s_6(f_0)$	$s_7(f_0)$
în	Xx	î	în	-	-	-	n	în	-	-	-	-	-
pereţii	x	p	pe	per	pere	pereţ	i	ii	ţii	eţii	reţii	ereţii	pereţii
boxei	x	b	bo	box	boxe	boxei	i	ei	xei	oxei	boxei	-	-
erau	x	e	er	era	erau	-	u	au	rau	erau	-	-	-
trei	x	t	tr	tre	trei	-	i	ei	rei	trei	-	-	-
orificii	x	o	or	ori	orif	orifi	i	ii	cii	icii	ficii	ificii	-

f_{-2}	l_{-2}	t_{-2}	f_{-1}	l_{-1}	m_{-1}	$s_1(m_{-2}) \oplus s_2(m_{-1})$	f_{+1}	$m_{train_{+1}}$
-	-	-	-	-	-	-	pereţii	{ Ncmpry}
-	-	-	în	în	Spsa	S	boxei	{}
în	în	Spsa	pereţii	perete	Ncmpry	S+N	erau	{Vmii3p}
pereţii	perete	Ncmpry	boxei	boxă	Ncfsoy	N+N	trei	{Mc-p-l}
boxei	boxă	Ncfsoy	erau	fi	Vmii3p	N+V	-	-
erau	fi	Vmii3p	trei	trei	Mc-p-l	V+M	-	-

Table 2: Features extracted for the morpho-tagging model from an example Romanian phrase: *În pereţii boxei erau trei orificii.*

On the morpho-tagging task, MBT is used without tuning algorithm parameters; the feature set closely mimics the one employed in **Morfette**: for known words we use two previous morpho-tags, two previous word-forms, the focus wordform, 7 suffixes and 5 prefixes of the focus wordform, one following wordform and its ambitag (set of tags it has been seen with in the training set); the same features minus the focus wordform are used for unknown words. With these settings MBT obtains somewhat lower morpho-tagging accuracy scores compared to **Morfette**: over 2% loss for Spanish and Romanian and over 6% for Polish.
The simple lemmatization algorithm described above performs quite well; nevertheless our method achieves for all languages over 60% relative error reduction rate for all words: 66.5% for Romanian, 65.4% for Spanish and 62% for Polish. As expected, a BASELINE model built this way shows poor performance on unseen words, in comparison to **Morfette**.

All words

	Morpho-tagging	Lemmatization	Joint
Romanian	94.49 (-2.34)	93.36 (-4.42)	90.21 (-5.87)
Spanish	93.13 (-2.27)	95.72 (-2.80)	90.70 (-4.32)
Polish	78.42 (-6.49)	88.29 (-7.26)	73.06 (-11.38)

Unseen words

	Morpho-tagging	Lemmatization	Joint
Romanian	72.01 (-14.67)	36.00 (-46.88)	23.10 (-55.40)
Spanish	59.40 (-16.31)	62.49 (-28.73)	31.70 (-40.14)
Polish	51.78 (-14.09)	20.82 (-60.29)	10.92 (-52.24)

Table 6: Evaluation results with BASELINE model with the FULL training set. Numbers in brackets compare accuracy with **Morfette** using the FULL training set

4. Error Analysis

We have performed detailed error analysis for morphological tagging and lemmatization for Spanish, Romanian and Polish. In this section we summarize the results of this analysis and suggest possible ways of dealing with some of the common errors our systems makes.

Errors in morphological tagging and lemmatization tend to co-occur: often an incorrectly assigned morphological category triggers lemmatization which is consistent with this category but incorrect given the gold morpho-tag. We will therefore discuss the issues related to both morphological tags and lemma-class tags jointly.

Named entities A common source of errors in Spanish and Romanian is failure to detect proper names (the tagset used in Polish does not have a separate tag for proper nouns). This results in the assignment of the wrong morphological tags and sometimes also the wrong lemma-class. For example in Spanish certain person or place names, such as *Reyes* or *Chiapas* have the plural suffix but, unlike for common nouns, their correct lemma-class should not delete it. Poor performance in this area is to be expected as our focus here is on learning morphological structure and not on detecting and classifying named entities. The only feature designed to capture some characteristics of those is $sp(F_0)$, the spelling pattern feature, which is clearly very rudimentary. In order to deal with named entities properly a dedicated module would be probably the best solution.

Suffix ambiguity A common phenomenon in all the three languages is suffix ambiguity, i.e. certain word endings can be indicative of more than one morphological category. In Spanish and Romanian, nouns and adjectives are difficult to distinguish based only on word endings and are sometimes mistagged and mislemmatized. This happen mostly in constructions with adjectives preceding nouns, which are rare and marked in comparison to adjectives post-modifying the noun.
In Romanian third person singular verbs in the *imperfect* tense have the same ending as nouns marked with a definite feminine article, and are also sometimes misclassified.[3]

Syncretism This is an especially frequent error type for Polish. Often different grammatical cases of the same lexical item have the same form, i.e. feminine genitive singular

[3]This affects only the written language as in speech those two forms differ in stress.

noun forms and feminine genitive plural forms, or masculine singular nominative and accusative.

There is sometimes genuine semantic ambiguity in the sentence but in many other cases, especially for number ambiguity, the correct morphological tag can be determined from context, but our system fails to do so. The determination of the right grammatical case is more difficult as it often involves non-local dependencies on the head verb or preposition and is unlikely to be solved completely by examining local context only.

Ambiguous function words Some high frequency function words are ambiguous: Spanish *que* (coordinating conjunction or relative pronoun), *se* (third person pronoun or impersonal pronoun); Romanian *a* (infinitive particle or a form of auxiliary *avea*, "have"), *lui* and *o* (article or pronoun); Polish *na* (locative or directional preposition). These distinctions are based on function rather than form and can be difficult to determine locally.

Annotation problems A nonnegligible number of errors in both morphological tagging and lemmatization are actually mistakes or inconsistencies in the training and test data. In the Polish dataset de-verbal nouns such as *działanie*, "doing", are sometimes tagged as nouns and sometimes as gerunds (where the corresponding lemma is the verb infinitive). There seems to be no consistent pattern to which tag is used when. Some Spanish plurals are assigned incorrect lemmas in the corpus.

Prefixal morphology Even though in the languages we examined inflectional morphology is almost exclusively suffixal, Polish offers one isolated but important exception. The superlative form of adjectives is formed by attaching the prefix *naj-* to the (already inflected) comparative form. Thus the comparative of *wysoki*, "tall", is *wyższy*, and the superlative is *najwyższy*. Lemma-classes are computed using the *shortest edit script* on reverse form and lemma, and this class induction method fails to generalize over word initial transformations. As a result, lemmas for superlatives are correct only in the case of very frequent words, and in general are not predicted correctly.

5. Conclusion

Morfette has two important features. Firstly, it is modular in the sense that the morphological-tagging and lemmatization models can use different features, can be trained separately, and even use different classifiers. Secondly, in spite of such modularity, the way our search algorithm combines morpho-tag and lemma-class conditional probabilities means that the two outputs of the two models are integrated at decoding time and their predictions are combined into an overall scoring over morpho-tag - lemma-class pair sequences.

From the evaluation and error analysis performed for three languages we have found that some error categories occur in all three languages; others are language and corpus specific. We suspect that the error classes which mostly affect unknown words could be dealt with successfully by (i) providing more training data, (ii) incorporating language-specific resources such as gazeteers or lexicons into our model. Other problems such as nominal/accusative syncretism or some ambiguous function words are more of a challenge, and although some improvement may be obtained by using more context and smarter features, it may be necessary to defer ambiguity resolution until a full syntactic structure is built.

Finally, the lemma-class induction mechanism is biased to dealing with suffixal morphology exclusively. We are currently experimenting with versions of this approach which feature a more linguistically accurate learning bias, and we expect this new version of **Morfette** to successfully deal with combined prefixal-suffixal morphological phenomena.

6. Acknowledgments

The authors gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for Grzegorz Chrupała and from DFG studentship in the International Research Training Group "Language Technology and Cognitive Systems" for Georgiana Dinu.

7. References

Alexandru Ceauşu. 2006. Maximum Entropy Tiered Tagging. In *Proceedings of ESSLLI 2006 Student Session*.

Grzegorz Chrupała. 2006. Simple data-driven context-sensitive lemmatization. *Procesamiento del Lenguaje Natural*, (37):121–127.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch, and Ko van der Sloot. 2007. MBT: Memory-Based Tagger, version 3.1, Reference Guide. ILK Technical Report Series 07-08.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: Lemmatizing unknown Slovene words. *Applied Artificial Intelligence*, 18:17–41.

Tomaž Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth Intl. Conf. on Language Resources and Evaluation*.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. In *Proceedings of the 17th international conference on Computational linguistics*, pages 483–490.

Jan Hajič. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 94–101.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gokhan Tür. 2002. Statistical Morphological Disambiguation for Agglutinative Languages. *Computers and the Humanities*, 36(4):381–410.

Chung-Hye Han and Martha Palmer. 2004. A morphological tagger for korean: Statistical tagging combined with

corpus-based morphological rule application. *Machine Translation*, 18(4):275–297.

M. Antonia Martí, Mariona Taulé, Lluís Márquez, and Manuel Bertran. 2007. CESS-ECE: A multilingual and multilevel annotated corpus. Available from: http://www.lsi.upc.edu/˜mbertran/cess-ece.

Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1):251–266, March.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dan Tufiş and Liviu Dragomirescu. 2004. Tiered Tagging Revisited. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC*, pages 39–42.

Dan Tufiş. 1999. Tiered tagging and combined language models classifiers. In *TSD '99: Proceedings of the Second International Workshop on Text, Speech and Dialogue*, pages 28–33.

Deniz Yuret and Ferhan Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 328–334.