SPMRL'13 Shared Task System: The CADIM Arabic Dependency Parser

Yuval Marton
Microsoft Corporation
City Center Plaza
Bellevue, WA, USA

Nizar Habash, Owen Rambow
CCLS
Columbia University
New York, NY, USA

Sarah Alkuhlani
CS Department
Columbia University
New York, NY, USA

cadim@ccls.columbia.edu

Abstract

We describe the submission from the Columbia Arabic & Dialect Modeling group (CADIM) for the Shared Task at the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL'2013). We participate in the Arabic Dependency parsing task for predicted POS tags and features. Our system is based on Marton et al. (2013).

1 Introduction

In this paper, we discuss the system that the Columbia Arabic & Dialect Modeling group (CADIM) submitted to the 2013 Shared Task on Parsing Morphologically Rich Languages (Seddah et al., 2013). We used a system for Arabic dependency parsing which we had previously developed, but retrained it on the training data splits used in this task. We only participated in the Arabic dependency parsing track, and in it, only optimized for predicted (non-gold) POS tags and features.

We first summarize our previous work (Section 2). We then discuss our submission and the results (Section 3).

2 Approach

In this section, we summarize Marton et al. (2013). We first present some background information on Arabic morphology and then discuss our methodology and main results. We present our best performing set of features, which we also use in our SPMRL'2013 submission.

2.1 Background

Morphology interacts with syntax in two ways: agreement and assignment. In *agreement*, there is coordination between the morphological features of two words in a sentence based on their syntactic configuration (e.g., subject-verb or noun-adjective agreement in GENDER and/or NUMBER). In *assignment*, specific morphological feature values are assigned in certain syntactic configurations (e.g., CASE assignment for the subject or direct object of a verb).

The choice of optimal linguistic features for a parser depends on three factors: relevance, redundancy and accuracy. A feature has **relevance** if it is useful in making an attachment (or labeling) decision. A particular feature may or may not be relevant to parsing. For example, the GENDER feature may help parse the Arabic phrase باب السيارة الجديد/الجديدة *bAb AlsyArħ Aljdyd/Aljdydħ*[1] 'door the-car the-new$_{masc.sg/fem.sg}$ [lit.]' using syntactic agreement: if *the-new* is masculine (*Aljdyd* الجديد), it should attach to the masculine *door* (*bAb* باب), resulting in the meaning 'the car's new door'; if *the-new* is feminine (*Aljdydħ* الجديدة), it should attach to the feminine *the-car* (*AlsyArħ* السيارة), resulting in 'the door of the new car'. In contrast, the ASPECT feature does

[1] Arabic orthographic transliteration is presented in the HSB scheme (Habash et al., 2007): (in alphabetical order)
ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
A b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional letters: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, pages 86–90,
Seattle, Washington, USA, 18 October 2013. ©2013 Association for Computational Linguistics

not constrain any syntactic decision.[2] Even if relevant, a feature may not necessarily contribute to optimal performance since it may be **redundant** with other features that surpass it in relevance. For example, the DET and STATE features alone both help parsing because they help identify the *idafa* construction (the modificiation of a nominal by a genitive noun phrase), but they are redundant with each other and the DET feature is more helpful since it also helps with adjectival modification of nouns. Finally, the **accuracy** of automatically predicting the feature values (ratio of correct predictions out of all predictions) of course affects the value of a feature on unseen text. Even if relevant and non-redundant, a feature may be hard to predict with sufficient accuracy by current technology, in which case it will be of little or no help for parsing, even if helpful when its gold values are provided. The CASE feature is very relevant and not redundant, but it cannot be predicted with high accuracy and overall it is not useful.

Different languages vary with respect to which features may be most helpful given various tradeoffs among these three factors. It has been shown previously that if the relevant morphological features in assignment configurations can be recognized well enough, then they contribute to parsing accuracy. For example, modeling CASE in Czech improves Czech parsing (Collins et al., 1999): CASE is relevant, not redundant, and can be predicted with sufficient accuracy. However, it had been more difficult showing that agreement morphology helps parsing, with negative results for dependency parsing in several languages (Nivre et al., 2008; Eryigit et al., 2008; Nivre, 2009). In contrast to these negative results, Marton et al. (2013) showed positive results for using agreement morphology for Arabic.

2.2 Methodology

In Marton et al. (2013), we investigated morphological features for dependency parsing of Modern Standard Arabic (MSA). The goal was to find a set of relevant, accurate and non-redundant features. We used both the MaltParser (Nivre, 2008) and the Easy-First Parser (Goldberg and Elhadad, 2010). Since the Easy-First Parser performed better, we use it in all experiments reported in this paper.

For MSA, the space of possible morphological features is quite large. We determined which morphological features help by performing a search through the feature space. In order to do this, we separated part-of-speech (POS) from the morphological features. We defined a core set of 12 POS features, and then explored combinations of morphological features in addition to this POS tagset. This core set of POS tags is similar to those proposed in cross-lingual work (Rambow et al., 2006; Petrov et al., 2012). We performed this search independently for Gold input features and predicted input features. We used our MADA+TOKAN system (Habash and Rambow, 2005; Habash et al., 2009; Habash et al., 2012) for the prediction. As the Easy-First Parser predicts links separately before labels, we first optimized for unlabeled attachment score, and then optimized the Easy-First Parser labeler for label score.

As had been found in previous results, assignment features, specifically CASE and STATE, are very helpful in MSA. However, in MSA this is true only under gold conditions: since CASE is rarely explicit in the typically undiacritized written MSA, it has a dismal accuracy rate, which makes it useless when used in machine-predicted (real, non-gold) condition. In contrast with previous results, we showed that agreement features are quite helpful in both gold and predicted conditions. This is likely a result of MSA having a rich agreement system, covering both verb-subject and noun-adjective relations.

Additionally, almost all work to date in MSA morphological analysis and part-of-speech (POS) tagging has concentrated on the morphemic form of the words. However, often the functional morphology (which is relevant to agreement, and relates to the meaning of the word) is at odds with the "surface" (form-based) morphology; a well-known example of this are the "broken" (irregular) plurals of nominals, which often have singular-form morphemes but are in fact plurals and show plural agreement if the referent is rational. In Marton et al. (2013), we showed that by modeling the functional morphology rather than the form-based morphology, we obtain a further increase in parsing performance

[2] For more information on Arabic morphology in the context of natural language processing see Habash (2010). For a detailed analysis of morpho-syntactic agreement, see Alkuhlani and Habash (2011).

Feature Type	Feature	Explanation
Part-of-speech	CORE12	12 tags for core parts-of-speech: verb, noun, adjective, adverb, proper noun, pronoun, preposition, conjunction, relative pronoun, particle, abbreviation, and punctuation
Inflectional features	DET	Presence of the determiner morpheme ال *Al*
	PERSON	1st, 2nd, or 3rd
	FN*N	Functional number: singular, dual, plural
	FN*G	Functional gender: masculine or feminine
Lexical features	FN*R	Rationality: rational, irrational, ambiguous, unknown or N/A
	LMM	Undiacritized lemma

Table 1: Features used in the CADIM submission with the Easy-First Parser (Goldberg and Elhadad, 2010).

Training Set	Test Set	LAS	UAS	LaS
5K (SPMRL'2013)	dev $\leq$ 70	81.7	84.7	92.7
All (SPMRL'2013)	dev $\leq$ 70	84.8	87.4	94.2
Marton et al. (2013)	test (old split) $\leq$ 70	81.7	84.6	92.8
5K (SPMRL'2013)	dev	81.1	84.2	92.7
All (SPMRL'2013)	dev	84.0	86.6	94.1
5K (SPMRL'2013)	test	80.5	83.5	92.7
All (SPMRL'2013)	test	83.2	85.8	93.9
Marton et al. (2013)	test (old split)	81.0	84.0	92.7

Table 2: Results of our system on Shared Task test data, Gold Tokenization, Predicted Morphological Tags; and for reference also on the data splits used in our previous work (Marton et al., 2013); "$\leq$ 70" refers to the test sentences with 70 or fewer words.

Training Set	Test Set	Labeled Tedeval Score	Unlabeled Tedeval Score
5K (SPMRL'2013)	test $\leq$ 70	86.4	89.9
All (SPMRL'2013)	test $\leq$ 70	87.8	90.8

Table 3: Results of our system on on Shared Task test data, Predicted Tokenization, Predicted Morphological Tags; "$\leq$ 70" refers to the test sentences with 70 or fewer words

(again, both when using gold and when using predicted POS and morphological features).

We also showed that for parsing with predicted POS and morphological features, training on a combination of gold and predicted POS and morphological feature values outperforms the alternative training scenarios.

2.3 Best Performing Feature Set

The best performing set of features on non-gold input, obtained in Marton et al. (2013), are shown in Table 1. The features are clustered into three types.

- First is part-of-speech, represented using a "core" 12-tag set.

- Second are the inflectional morphological features: determiner clitic, person and functional gender and number.

- Third are the rationality (humanness) feature, which participates in morphosyntactic agreement in Arabic (Alkuhlani and Habash, 2011), and a form of the lemma, which abstract over all inflectional morphology.

For the training corpus, we use a combination of the gold and predicted features.

3 Our Submission

3.1 Data Preparation

The data split used in the shared task is different from the data split we used in (Marton et al., 2013), so we retrained our models on the new splits (Diab et al., 2013). The data released for the Shared Task showed inconsistent availability of lemmas across gold and predicted input, so we used the ALMOR analyzer (Habash, 2007) with the SAMA databases (Graff et al., 2009) to determine a lemma given the word form and the provided (gold or predicted) POS tags. In addition to the lemmas, the ALMOR analyzer also provides morphological features in the feature-value representation our approach requires. Finally, we ran our existing converter (Alkuhlani and Habash, 2012) over this representation to obtain functional number and gender, as well as the rationality feature.[3] For simplicity reasons, we used the MLE:W2+CATiB model (Alkuhlani and Habash, 2012), which was the best performing model on seen words, as opposed to the combination system that used a syntactic component with better results on unseen words. We did not perform Alif or Ya normalization on the data.

We trained two models: one on 5,000 sentences of training data and one on the entire training data.

3.2 Results

Our performance in the Shared Task for Arabic Dependency, Gold Tokenization, Predicted Tags, is shown in Table 2. Our performance in the Shared Task for Arabic Dependency, Predicted Tokenization, Predicted Tags, is shown in Table 3. For predicted tokenization, only the IMS/Szeged system which uses system combination (Run 2) outperformed our parser on all measures; our parser performed better than all other single-parser systems. For gold tokenization, our system is the second best single-parser system after the IMS/Szeged single system (Run 1). For gold tokenization and predicted morphology (Table 2), we also give the performance reported in our previous work (Marton et al., 2013). The increase over the previously reported work may simply be due to the different split for training and test, but it may also be due to improvements to the functional feature prediction (Alkuhlani and Habash, 2012), and the predicted features provided by the Shared Task organizers.

[3] The functional feature generator of (Alkuhlani and Habash, 2012) was trained on a different training set from the parser, but the functional feature generator was not trained on any of the test corpus for the Shared Task.

References

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, Oregon, USA.

Sarah Alkuhlani and Nizar Habash. 2012. Identifying broken plurals, irregular gender, and rationality in Arabic text. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–685. Association for Computational Linguistics.

Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 505–512, College Park, Maryland, USA, June.

Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic Treebanks and Associated Corpora: Data Divisions Manual. Technical Report CCLS-13-02, Center for Computational Learning Systems, Columbia University.

Gülsen Eryigit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Proceedings of Human Language Technology (HLT): the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 742–750, Los Angeles, California.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

Nizar Habash, Owen Rambow, and Ryan Roth. 2012. MADA+TOKAN Manual. Technical report, Technical Report CCLS-12-01, Columbia University.

Nizar Habash. 2007. Arabic Morphological Representations for Machine Translation. In Antal van den Bosch and Abdelhadi Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Yuval Marton, Nizar Habash, and Owen Rambow. 2013. Dependency parsing of Modern Standard Arabic with lexical and inflectional features. *Computational Linguistics*, 39(1).

Joakim Nivre, Igor M. Boguslavsky, and Leonid K. Iomdin. 2008. Parsing the SynTagRus Treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 641–648.

Joakim Nivre. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 34(4).

Joakim Nivre. 2009. Parsing Indian languages with MaltParser. In *Proceedings of the ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, pages 12–18.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, May.

Owen Rambow, Bonnie Dorr, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith J. Miller, Teruko Mitamura, Florence Reeder, and Siddharthan Advaith. 2006. Parallel syntactic annotation of multiple languages. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2013. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*, Seattle, WA.