NP Subject Detection in Verb-Initial Arabic Clauses

Spence Green, Conal Sathi, and Christopher D. Manning
Computer Science Department
Stanford University
Stanford, CA 94305
{**spenceg,csathi,manning**}**@stanford.edu**

Abstract

Phrase re-ordering is a well-known obstacle to robust machine translation for language pairs with significantly different word orderings. For Arabic-English, two languages that usually differ in the ordering of subject and verb, the subject and its modifiers must be accurately moved to produce a grammatical translation. This operation requires more than base phrase chunking and often defies current phrase-based statistical decoders. We present a conditional random field sequence classifier that detects the full scope of Arabic noun phrase subjects in verb-initial clauses at the $F_{\beta=1}$ 61.3% level, a 5.0% absolute improvement over a statistical parser baseline. We suggest methods for integrating the classifier output with a statistical decoder and present preliminary machine translation results.

1 Introduction

Arabic to English translation often requires multiple, significant phrase re-orderings. In particular, the verb-initial clauses that are a characteristic feature of Arabic must be inverted for Subject-Verb-Object (SVO) target languages like English. To demonstrate the strain this requirement places on phrase-based statistical decoders, consider the VOS Arabic example in Figure 1. The noun phrase (NP) subject is the recursive Arabic annexation structure **الإضافة** *iDafa* in which the rightmost noun is modified by a chain of nouns and adjectives. The decoder must accurately identify the full NP subject and move it four positions to the left under the following conditions:

- the length of the NP subject approaches the maximum phrase length used in translation models[1]
- the required horizontal movement nears the distortion limit commonly used in phrase-based decoders
- each recursive level in the NP is grammatical

The last condition causes the language model to license different hypotheses that are grammatical but semantically inconsistent with the source language (e.g., *Followers waited for all of the Christian and Islamic sects*). This is not a rare example. If we take the Penn Arabic Treebank (ATB) (Maamouri et al., 2004) as a guide, then over 25% of the NP subjects in Arabic verb-initial clauses are of length five or greater (Table 1).

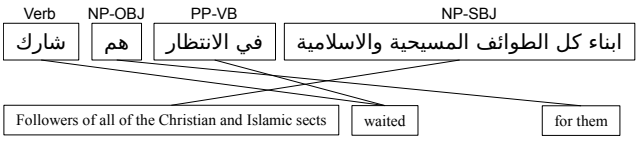


Figure 1: A VOS sentence from the ATB. The Arabic phrases read right-to-left, but we have ordered the sentence from left-to-right in order to clearly illustrate the re-ordering problem.

Until the feature-poor statistical MT models currently in use are improved (Avramidis and Koehn, 2008), distortion limits will be necessary to both make decoding tractable (Knight, 1999) and to improve translation quality. However, if the scope of Arabic NP subjects could be accurately identified,

[1]Our best Arabic-English system uses a maximum phrase length of 5 and a distortion limit of 4.

Length	Frequency
1	34.42%
2	21.90%
3	10.28%
4	6.86%
5	5.68%
6	3.78%
7-10	8.62%
11-30	7.39%
31-131	1.08%

Table 1: ATB frequencies for *maximal* NP subjects in verb-initial clauses. The average subject length is 4.28 words with a maximum observed length of 131 words.

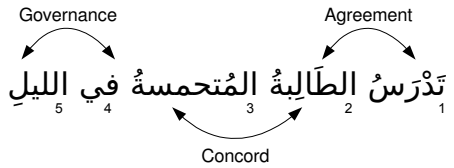
then a component could be added to MT systems to encourage particular re-orderings that would be otherwise unlikely under the conditions shown in Figure 1.

We present a conditional random field (CRF) sequence classifier that detects the full scope of NP subjects in verb-initial Arabic clauses. The assignment of grammatical relations to sentences has traditionally required a parser, although the popular Arabic parsers of Bikel (2004) and Klein and Manning (2002) do not support grammatical relations by default. Not only does our classifier greatly exceed the performance of these two statistical parsers, but it also processes MT test sets in seconds. The best feature set finds subjects at the $F_{\beta=1}$ 61.3% level, a 5.0% absolute improvement over the best parser baseline. We analyze current classifier results, suggest strategies for integrating the classifier output with a phrase-based decoder, and provide a preliminary MT evaluation.

2 Background

2.1 Linguistic Motivation

Schmid and Rooth (2001) operationalize the syntactic notion of governance for the detection of grammatical relations. We extend this idea to Arabic by capitalizing on its relatively rich set of syntactic *dependency relations* (Ryding, 2005). In addition to *governance*—which is the phenomenon in which certain words cause dependents to inflect in specific ways—we are also concerned with *concord* and *agreement*. Concord refers to matching between nouns and dependents (e.g., adjectives) for features such as definiteness and case. When compatibility



Gloss: **studies the student the enthusiastic in the night*

Figure 2: Example of Arabic dependency relations. The verb (1) is in **agreement** with the noun (2) in both gender and number. The **concord** dependency requires the adjective (3) to have a feminine affix to match the gender of the noun. Finally, the preposition (4) **governs** its noun (5), causing it to inflect in the genitive case.

between verbs and subjects is in question, however, agreement may be checked for other features such as gender and number (Figure 2). In Arabic, agreement may be either rich (matching for gender and number) or poor (in which only gender matches).

A final syntactic feature of Arabic related to these dependencies is *case*. Case is explicitly marked in Arabic, typically by a short vowel suffix. Modification to a long vowel suffix can also indicate case. There are three cases in Arabic: nominative (which almost always indicates a subject, especially in verb-initial configurations), accusative, and genitive.

Along with these surface features, a particular account of syntactic movement in the deep structure also influences our approach. Fassi Fehri (1993) argues in favor of the SVO X-bar schema in Figure 3 as the canonical phrase structure instantiated by the Arabic grammar. If this is so, then a transformation rule is required to explain the movement of the V node (the verb) to the I position (ahead of the NP subject) at the surface.[2] Fassi Fehri (1993) claims that poor agreement, which prevents the NP subject from raising to I, is precisely that rule. Because the verb "protects" the subject from other governors, V to I raising also enables nominative declension of the NP subject. This theory appears to account for the admissibility of other case markings for SVO subjects, while VSO and VOS subjects are always nominative.

[2]Discussions of word ordering have been the source of considerable controversy among Arabic grammarians. We neither posit a novel claim nor take a definite side, but simply describe one theory that has helped us understand the task.

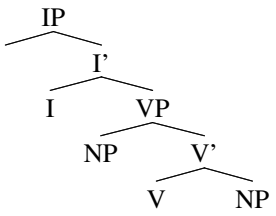


Figure 3: Canonical Arabic phrase structure of Fassi Fehri (1993).

From this description, a method for detecting subjects seems obvious: why not develop a set of rules to match NPs in the nominative case? Here convention subverts our project. The Arabic MT data that we use is usually unvocalized: the short vowels and other diacritics, including case markings, are dropped.[3] From this perspective, the task of subject detection in verb-initial Arabic clauses is better formulated as an attempt to recover information omitted by convention.

Anecdotal evidence suggests that this is a viable objective. When case is not explicitly marked, as is common in spoken Arabic, native speakers often choose SVO and rich agreement to reduce ambiguity. In written Arabic, however, there is a stylistic bias toward VSO *without* case markings. Consequently, writers often choose "pragmatically-neutral contexts", or structures that require comparatively little in terms of interpretive capacity. Our work thus turns on the hypothesis that in MT data, other surface–level dependency relations in the unvocalized text are sufficient to identify subjects.

2.2 Conditional Random Fields

We use a conditional random field (CRF) sequence model (Lafferty et al., 2001) for two reasons. First, CRF classifiers have a relatively high modeling capacity and can thus accommodate the high number of overlapping features that we want to encode. In our task, CRFs also provide a more convenient method for specifying relationships between words than do parsers. In canonical form, linear-chain CRFs are defined as a globally normalized product of real-valued state functions s and transition functions t, where $\mathbf{x}$ is the sequence of observations and $\mathbf{y}$ is the set of labels for those observations:

$$s(y_i|\mathbf{x}) = \exp(\sum_i \lambda_i f_i(y_i, \mathbf{x})) \quad (1)$$

$$t(y_i, y_{i-1}|\mathbf{x}) = \exp(\sum_j \lambda_j g_j(y_{i-1}, y_i, \mathbf{x})) \quad (2)$$

We then find the assignment of labels to observations that maximizes the probability of the sequence (where $Z(\mathbf{x})$ is an appropriate normalizing constant):

$$p(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \prod_i t(y_i, y_{i-1}, x_i) s(y_i, x_i) \quad (3)$$

CRFs can generalize to higher-order cliques, a capability that we leverage. Wallach (2004) discusses this and other aspects of CRFs in greater detail.

3 Design

3.1 Data Annotation

We treat subject detection as a supervised learning task. For training data, we use the first three parts of the ATB.[4] Subjects in the ATB are labeled using the "dashtags" in Table 2 (Maamouri et al., 2009). We relate these tag names to the previous linguistic discussion by identifying five categories of subjects present in the data.

Subject inside VP This category accounts for both VSO and VOS configurations, the subjects of which are marked by NP-SBJ. We mark all non-trace NP-SBJ spans in the classifier training data. Since these subjects require re-ordering during translation to English, we tailor our feature set to them.

Null Subjects Null subjects are one of the most vexing issues for Arabic-English MT. They result from pro-drop clauses in which there is no lexical subject. Instead, inflection on the verb indicates the gender, number, and person of the dropped pronominal subject. In this case, the direct object often appears adjacent to the verb and we do not want to perform re-ordering. The ATB marks null subjects with an NP-SBJ trace inside the VP, but this data is obviously not present in MT input. The CRF classifier does not explicitly indicate null subjects. Instead, it

[3] Automatic vocalization tools do exist, but we have not experimented with them.

[4] LDC A-E catalog numbers: LDC2008E61 (ATBp1v4), LDC2008E62 (ATBp2v3), and LDC2008E22 (ATBp3v3.1).

is designed to avoid marking a subject in such scenarios.

Topicalized Subjects Pre-verbal subjects (SVO) are marked with NP-TPC. In the presence of a topicalized subject, an NP-SBJ trace is always included inside the sister VP (Maamouri et al., 2009). These clauses do not require re-ordering, so we do not mark them during training. We also remove the NP-SBJ trace during pre-processing.

Clausal Subjects Certain Arabic words such as the pseudo-verbs (إن وأخواتها) take clausal subjects, which are marked with S-NOM-SBJ, SBAR-NOM-SBJ, or SBAR-SBJ. An example from the ATB illustrates how we should handle clausal subjects:

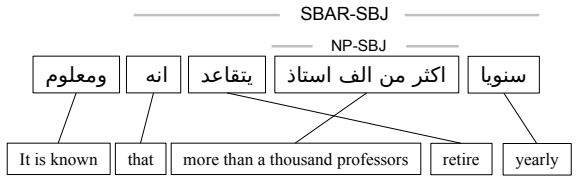


Figure 4: NP-SBJ requires re-ordering around the verb, while SBAR-SBJ should not be re-ordered. As in Figure 1, the Arabic phrases read right-to-left, but the sentence is arranged left-to-right.

Clearly the span covered by SBAR-SBJ is not re-ordered in the translation. Instead, we want to swap the NP subject *more than a thousand professors* with the verb *retire*. We thus do not mark clausal subjects in training, focusing instead on NP subjects inside clausal subjects.

Verbs of being/becoming (أخوات كان) Verbs of being and becoming merit special attention in Arabic. In equational sentences (nominal sentences in traditional Arabic grammars), they are omitted, thus resulting in an SVO configuration without an explicit verb. In the verb-initial case, the NP subject usually appears immediately after the verb, but inversion of the subject and an NP predicate is possible. Moreover, the NP subject is often pro-drop, so it is not explicit. As before, we include the NP-SBJ constituents in training, but we must carefully design features to handle the pro-drop and inverted cases.

To prepare the classifier training data, we use

Function Tag	Description
NP-SBJ	Subject inside VP, including null subjects
S-NOM-SBJ	Clausal subjects
SBAR-NOM-SBJ	Clausal subjects
SBAR-SBJ	Clausal subjects
S-SBJ	Co-extensive with some quotations (rare)
NP-TPC	Topicalized subjects (SVO ordering)

Table 2: Of the subject-related functional tags in the ATB, we only include NP-SBJ dominated by VP in training.

Tregex expressions to identify NP-SBJ constituents in verb-initial clauses (Levy and Andrew, 2006). Using the Bies mappings provided with the ATB, we convert the pre-terminal morphological analyses to shortened part-of-speech (POS) tags. We augment shortened tags for definite nouns with "DT", which improves performance. We then linearize the parse trees and label each word according to the classes in Table 3. A reserved symbol separates sentences. Finally, delimiters are attached to the beginning and end of each word so that initial, medial, and final character n-grams may be identified.

3.2 Morphological Information

In addition to the POS tags and words, we add morphological data to the classifier input. We run MADA 2.12, a morphological pipeline, on each input sentence (Habash and Rambow, 2005). MADA first uses a morphological analyzer to generate an n-best list of analyses for each word. It then re-ranks the n-best lists using a weighted set of support vector machine (SVM) classifiers. We use the stock classifier set—which includes number and person—plus the gender classifier, which we give a 0.02 weight. For verbs, we also retrieve the verb stem. Two sets of data result from this procedure: the output of the classifiers, and the top-ranked morphological analyses. We take the verb stem from the morphological analyses and all other features from the SVM output. Including the words and POS tags, we have an observation set $\mathbf{x} = \langle x_1, x_2, x_3, ... \rangle$, where each observation x_i is a tuple consisting of (word, POS tag, gender, number, person, case, verb stem).

Class	Description
NP_SUBJ_START	Beginning of a subject of length 2 or more
NP_SUBJ_IN	Between the START and END labels
NP_SUBJ_END	Last word in a subject of length 2 or more
NP_SUBJ_SINGLE	Singleton class for 1-word subjects
NP_NOT_SUBJ	Non-subject noun phrases
VERB	Verb types
OTHER	All other constituents

Table 3: Label inventory for the CRF classifier.

3.3 Labels

The classifier assigns one of seven different labels from the set $\mathbf{y}$ to each token (Table 3). The label set is derived primarily from observations about the sequence model. First, we note that all subjects are NPs, but not all NPs are subjects. We therefore define specific subject labels, and confine all other NPs to a single negative class. The START, IN, and END subject labels help the model learn strong higher-order clique potentials (i.e., the model learns that IN follows START, END follows IN, and so on). We add a singleton subject type for single-word subjects. This is particularly effective for pronominal subjects. To these subject labels we add a VERB class so that the model learns that VERB usually precedes START. Finally, we assign all other constituents to an OTHER category.

3.4 Classifier Features

Table 4 lists the CRF feature set. Space limitations prevent expanded commentary, but we provide brief feature descriptions. The strongest features, as indicated by the feature values learned during training, are **pos-tags**, **pp-vb-pairs**, **temporal-nn**, **inna**, and **path**. Experiments on the development set led to the final model, which uses features 2 and 7-19. This model is remarkable in that it does not use **word**, a customary feature in classical sequence model tasks like named entity recognition. For subject detection, we found that **word** creates significant overfitting. It is also worth mentioning that BAMA 2.0, the morphological analyzer used by MADA, can emit nominative case markings. Given the preceding linguistic discussion, we expected noise in this data. Experiments confirmed this intuition, so the final model does not use **nom-case**.

	Feature	Description
1	**word**	The current word in the sequence
2	**pos-tags**	POS tags for a configurable window of observations
3	**collapse-tags**	Collapse noun tags to NN and verb tags to VB
4	**word-prefix**	Look for determiner ال *Al*
5	**word-suffix**	Look for feminine ة *p* and accusative ًا *A* suffixes
6	**nom-case**	Nominative case in the morphological data
7	**acc-case**	Accusative suffix ًا *A* on consecutive adjectives and indefinite nouns
8	**concord**	Gender concord for consecutive nouns and adjectives
9	**conj-break**	Conjunctions preceeding non-nouns
10	**agreement**	Establish agreement between nouns and verbs
11	**aux-pairs**	Mark verbs of becoming (أخوات كان) and arguments
12	**inna**	Mark pseudo-verbs (إن وأخواتها) and arguments
13	**qp-matching**	Mark close quotes and parentheses
14	**pp-vb-pairs**	Associate preposition with stem of most recent verb
15	**temporal-nn**	Mark temporal nouns (days, months, etc.) and modifiers
16	**inna-pp-attach**	Specify noun attachment for PPs near pseudo-verbs
17	**path**	Adapted from Gildea and Jurafsky (2002)
18	**annexing**	الإضافة *iDafa* POS patterns
19	**vb-stem**	Observation has a verb stem

Table 4: Subject classifier features and descriptions.

We choose features that bias the classifier toward high precision. This decision is motivated by experience with integrating other syntactic MT features into phrase-based decoders. As a general rule, if a classification decision cannot be made with high confidence, then it is best to abstain from influencing decoding.

4 Evaluation

4.1 Subject Detection

We implement the CRF classifier using the publicly available package of Finkel et al. (2005) and modify the Arabic parsers of both Klein and Manning (2002) and Bikel (2004) to train on trees marked with subject-inside-VP constituents. We divide the ATB into training, development, and test sets using the split of Chiang et al. (2006).[5] To make the comparison fair, we use this split along with common orthographic normalization rules for both the classifier and parser experiments. We pre-tag the test set

[5]The original split contained the duplicate document ANN20021115.0092, which has been removed from the ATB.

for Bikel (2004) using the POS tagger of Toutanova et al. (2003), a practice that slightly enhances performance.[6]

The classifier training set is linearized and labeled according to the conventions described previously. We run MADA and the POS tagger of Toutanova et al. (2003) on the classifier test set instead of including the gold morphological analyses and POS tags from the ATB. This procedure replicates the MT test environment.

Table 5 lists results for both the parsers and two CRF models. We score precision, recall, and $F_{\beta=1}$ for *contiguous* subjects, i.e., credit for a classification is only awarded for identification of the full NP subject scope. Although the classifier is designed to identify subjects, it indirectly indicates verb-initial sentences by the absence of a labeled NP subject prior to the first verb in the sentence. In the same manner, it identifies equational sentences by omitting an NP subject label. Using these metrics, the best feature set finds verb-initial sentences with 98.1% accuracy.

4.2 Machine Translation

Our MT system uses a re-implementation of the Moses decoder (Koehn et al., 2007) with the same standard features: four translation features (phrase-based translation probabilities and lexically-weighted probabilities), word penalty, phrase penalty, linear distortion, and language model score.

The training set consists of 19.5M English words and 18.7M Arabic words originating from parallel news data released by the LDC. We create word alignments using the Berkeley Aligner (Liang et al., 2006) and perform symmetrization with the grow-diag heuristic.

We build a 5-gram language model from the Xinhua and AFP sections of the Gigaword corpus (LDC2007T40), in addition to the target side of the training data. We manually remove Gigaword documents that were released during periods that overlapped with the development and test sets. The language model is smoothed with the modified Kneser-Ney algorithm, retaining only trigrams, 4-grams, and 5-grams that occurred two, three, and three times, respectively, in the training data.

[6]When trained and tested on the same ATB split, the POS tagger achieves 96.4% accuracy.

System	Features	P	R	$F_{\beta=1}$
(Bikel, 2004)		50.9	59.7	55.0
(Klein and Manning, 2002)		55.2	57.5	56.3
CRF BASELINE	1,2	57.4	49.1	52.9
CRF BEST	2,7-19	**65.9**	**57.3**	**61.3**

Table 5: Test set performance of two CRF models v. statistical parser baselines. The feature indices correspond to Table 4.

BLEU			
	MT04 (dev)	MT03	MT05
BASELINE	48.69	52.63	53.57
BASELINE+SUBJ	48.66 (−0.03)	52.61 (−0.02)	53.43 (−0.14)
Translation Error Rate (TER)			
	MT04 (dev)	MT03	MT05
BASELINE	42.02	40.30	39.48
BASELINE+SUBJ	42.05 (+0.03)	40.51 (+0.21)	39.56 (+0.08)

Table 6: MT experimental results evaluated with the BLEU (Papineni et al., 2001) and TER (Snover et al., 2006) metrics.

The output of the CRF classifier is incorporated into the decoder using a simple model. We positively weight phrase pairs that fully overlap with the subject and penalize partial overlaps with a score inversely proportional to the size of the overlap. The feature therefore prefers hypotheses in which the subject is translated as a contiguous block. Hypotheses that split the subject—by inserting a verb inside the subject, for example—receive a negative score. A single feature weight for this model is set during MERT (Och, 2003). Table 6 shows results using this feature design.

5 Discussion

We have shown a considerable improvement in subject detection performance and, for completeness, have presented preliminary MT results. In this section, we analyze current sources of error and identify areas for improvement.

Subject/Object Boundary and PP Attachment We have specified a set of strong features that indicate the beginning of a subject, but have yet to discover a robust way to find the last word in a subject. "Catastrophic" chains of false positives can thus occur. In these scenarios, the classifier properly detects the beginning of a subject, but fails to identify its other boundary. It classifies every word as a subject until it encounters the sentence delimiter.

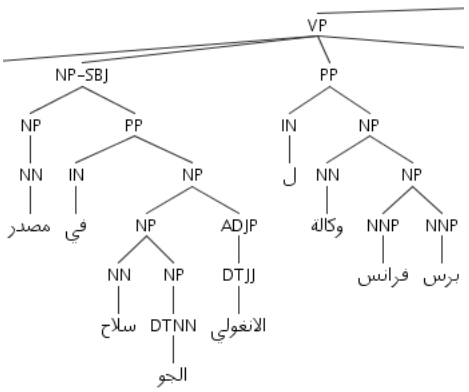


Figure 5: An example of the PP attachment problem. The classifier does not correctly identify the NP subject final boundary and thus includes the second of the two adjacent PPs in the subject.

Prepositional phrase (PP) attachment, a common problem in parsing, creates ambiguity at subject boundaries (Figure 5). The **path** feature does help in this regard, but it is not a comprehensive solution.

"Recursive" NP subjects Our experiments mark NP subjects that are not contained in other NP subjects (so-called *maximal* subjects). A non-trivial number of instances in the ATB contain NP subjects with internal NP subjects. This phenomenon is also present in English, as shown by this example:

Whether Peter is guilty and which people helped him are the issues.

Both *Peter* and *people* are the subjects of their respective clauses, while both are part of the main clause NP subject. Given that the accuracy of the classifier degrades with subject length, performance could be improved by only labeling NP subjects up to a given length (e.g., the phrase-limit used in a translation model). In this example, it is likely that the classifier could detect the two smaller clausal subjects, but would not identify the full scope of the main clause subject. In some MT settings, this may be a preferable strategy.

Machine Translation We have not yet obtained substantial gains in MT performance, but our initial experiments have revealed promising directions for future work. The key difference in experimental parameters between BASELINE and BASELINE+SUBJ is the distortion limit (recall that this parameter governs horizontal movement by words and phrases). When testing the subject feature, we initially set the linear distortion to four, above which performance degrades in our baseline system. We noticed a decrease in performance across all test sets: the feature had a negative effect relative to the baseline. We then set the distortion limit to five and ran MERT again. This time, the subject feature received a relatively high weight and performance was competitive with the baseline. These experiments suggest that even a simple realization of the feature does have useful discriminative capacity in that it encourages principled re-orderings as the distortion limit increases. We speculate that more substantial improvements could be realized with a feature design that utilizes word alignments. This investigation is left to future work.

6 Prior Work

Two groups of literature that are immediately relevant to our work investigate the assignment of grammatical relations to English text. The first group employs supervised and unsupervised learning techniques *during* parsing. Carroll and Briscoe (2002), who use a semi-lexicalized LR parser augmented with the governor annotation algorithm of Schmid and Rooth (2001), are representative of this group. In particular, they observe that certain tasks—of which ours is clearly a member—only benefit when a grammatical relation is marked with a high degree of confidence. They thus evaluate various thresholding techniques to boost precision to 90% from a baseline of 75%. No specific results for subjects are provided.

The other group of work describes the assignment of functional tags to parsed text as a post-processing step. Blaheta and Charniak (2000) use feature trees to recover semantic roles and other information after parsing. For grammatical relations—the category that includes subjects—they show an $F_{\beta=1}$ of 95.65%, although this figure excludes constituents that were incorrectly parsed (11% of the test set).

To our knowledge, no extant studies address the assignment of grammatical relations to Arabic text. Diab (2007) describes an SVM-based base phrase chunker that achieves an $F_{\beta=1}$ of 94.92% for base NPs, but subjects are not always co-extensive with

base (non-recursive) NPs. This is especially true of Arabic, in which the الإضافة *iDafa* construct, a type of recursive NP, is a characteristic feature. Moreover, we are unaware of any prior work that uses a CRF classifier to identify subjects.

7 Conclusion

We have presented a sequence classifier that detects NP subjects in verb-initial Arabic clauses with greater accuracy than current statistical parsers. Using a simple decoder integration technique, we have shown that knowledge of subject spans does allow more possibilities for accurate phrase re-ordering. In future experiments, we will use word alignments to improve the decoder feature.

Acknowledgments

We thank Daniel Cer, Jenny Finkel, and Michel Galley for helpful conversations, and Zavain Dar for his contribution to an earlier version of this work. The first author is supported by a National Defense Science and Engineering Graduate (NDSEG) fellowship. This paper is based on work supported in part by the Defense Advanced Research Projects Agency through IBM. The content does not necessarily reflect the views of the U.S. Government, and no official endorsement should be inferred.

References

E Avramidis and P Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proc. of ACL*.

DM Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

D Blaheta and E Charniak. 2000. Assigning function tags to parsed text. In *Proc. of NAACL*.

J Carroll and T Briscoe. 2002. High precision extraction of grammatical relations. In *Proc. of COLING*.

D Chiang, M Diab, N Habash, O Rambow, and S Shareef. 2006. Parsing Arabic dialects. In *Proc. of EACL*.

M Diab. 2007. Improved Arabic base phrase chunking with a new enriched POS tag set. In *Proc. of the 5th Workshop on Important Unresolved Matters*.

A Fassi Fehri. 1993. *Issues in the structure of Arabic clauses and words*. Kluwer Academic Publishers.

J Finkel, T Grenager, and CD Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. of ACL*.

D Gildea and D Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

N Habash and O Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proc. of ACL*.

D Klein and CD Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Cambridge, MA. MIT Press.

K Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4).

P Koehn, H Hoang, A Birch, C Callison-Burch, M Federico, N Bertoldi, B Cowan, W Shen, C Moran, R Zens, C Dyer, O Bojar, A Constantin, and E Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL, Demonstration Session*.

J Lafferty, A McCallum, and F Pereira. 2001. Conditional random fields: Probablistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

R Levy and G Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proc. of LREC*.

P Liang, B Taskar, and D Klein. 2006. Alignment by agreement. In *Proc. of NAACL*.

M Maamouri, A Bies, T Buckwalter, and W Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

M Maamouri, A Bies, S Krouna, F Gaddeche, and B Bouziri. 2009. Penn Arabic Treebank guidelines v4.8. Technical report, Linguistic Data Consortium, University of Pennsylvania.

FJ Och. 2003. Minimum error rate training for statistical machine translation. In *Proc. of ACL*.

K Papineni, S Roukos, T Ward, and W-J Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

K Ryding. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

H Schmid and M Rooth. 2001. Parse forest computation of expected governors. In *Proc. of ACL*.

M Snover, B Dorr, R Schwartz, L Micciulla, and J Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*.

K Toutanova, D Klein, CD Manning, and Y Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL*.

H Wallach. 2004. Conditional random fields: An introduction. Technical report, Department of Computer and Information Science, University of Pennsylvania.