Resumptive Pronoun Detection for Modern Standard Arabic to English MT

Stephen Tratz[*] **Clare Voss**[*] **Jamal Laoudi**[†]
[*]Army Research Laboratory, Adelphi, MD 20783
[†]Advanced Resource Technologies, Inc. Alexandria, VA 22314
{stephen.c.tratz.civ,clare.r.voss.civ,jamal.laoudi.ctr}@mail.mil

Abstract

Many languages, including Modern Standard Arabic (MSA), insert resumptive pronouns in relative clauses, whereas many others, such as English, do not, using empty categories instead. This discrepancy is a source of difficulty when translating between these languages because there are words in one language that correspond to empty categories in the other, and these words must either be inserted or deleted—depending on translation direction. In this paper, we first examine challenges presented by resumptive pronouns in MSA-English translations and review resumptive pronoun translations generated by a popular online MSA-English MT engine. We then present what is, to the best of our knowledge, the first system for automatic identification of resumptive pronouns. The system achieves 91.9 F1 and 77.8 F1 on Arabic Treebank data when using gold standard parses and automatic parses, respectively.

1 Introduction

One of the challenges for modern machine translation (MT) is the need to systematically insert or delete information that is overtly expressed in only one of the languages in order to maintain intelligibility and/or fluency. For example, word alignment between pro-drop and non-pro-drop languages can be negatively impacted by the systematic dropping of pronouns in only one of the languages (Xiang et al., 2013). A similar type of linguistic phenomenon of great interest to linguists that has not yet received significant attention in MT research is the mismatch between languages in their usage of resumptive pronouns. Some languages, such as Modern Standard Arabic (MSA), require the insertion of resumptive pronouns in many relative clauses, whereas other languages, including English, rarely permit them. An example of an MSA sentence is given below, with its English gloss showing the resumptive pronoun in bold, its reference translation (RT), and an MT system output where the roles of *patient* and *doctor* are incorrectly reversed:

رأيت المريض الذي أنقذته الطبيبة

Gloss: *I.saw the.patient who rescued.***him** *the.doctor.*
RT: *I saw the patient whom the doctor rescued.*
MT: *I saw a patient who rescued the doctor.*

In this paper, we examine translations produced by a popular online translation system for MSA resumptive pronouns occurring in several different syntactic positions to gain insight into the types of errors generated by current MT engines. In a test suite of 300 MSA sentences with resumptive pronouns, over 30% of the relative clauses with resumptive pronouns were translated inaccurately. We then present an automatic classifier that we built for identifying MSA resumptive pronouns and the results obtained from using it in experiments with the Arabic Treebank (Maamouri et al., 2004; Maamouri and Bies, 2004). The system achieves 91.9 F1 and 77.8 F1 on Arabic Treebank data when using gold standard parses and automatic parses, respectively. To the best of our knowledge, this is the first attempt to automatically identify resumptive pronouns in any language.

2 Relevant MSA Linguistics

MSA and English relative clauses differ in structure, with one of the most prominent differences being in regard to resumptive pronouns. Resumptive pronouns are required in many MSA relative clauses but are almost never grammatical in English. In MSA, like English, if the external

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 42–47,
Gothenburg, Sweden, April 27, 2014. ©2014 Association for Computational Linguistics

	Arabic (...أعرف)	Gloss (I know...)	English RT (I know...)	MT Output (I know...)
1a	السيدة التي تبتسم كثيرا	the+lady who$_i$ ϵ_i smiles a_lot	the lady who$_i$ ϵ_i smiles a lot	the lady who smiles a lot
1b	سيدة تبتسم كثيرا	lady ω_i smiles ϵ_i a_lot	a lady who$_i$ ϵ_i smiles a lot	a lot lady smiling
1c	من يبتسم كثيرا	who$_i$ smiles ϵ_i a_lot	who$_i$ ϵ_i smiles a lot	a lot of smiles
2a	الشركة التي مولها الرجل	the+company that$_i$ financed+it$_i$ the+man	the company that$_i$ the man financed ϵ_i	the company that financed the man
2b	شركة مولها الرجل	company ω_i financed+it$_i$ the+man	a company ω_i the man financed ϵ_i	a company funded by the man
2c	ما موله الرجل	what$_i$ financed+it$_i$ the+man	what$_i$ the man financed ϵ_i	what the man-funded
3a	الولد الذي تكلمت الفتاة معه	the+boy whom$_i$ talked the+girl with+him$_i$	the boy whom$_i$ the girl talked with ϵ_i	the boy who spoke with the girl
3b	ولدا تكلمت الفتاة معه	boy ω_i talked the+girl with+him$_i$	a boy ω_i the girl talked with ϵ_i	the girl was born I spoke with him
3c	مع من تكلمت الفتاة	[with whom]$_i$ talked the+girl ϵ_i	[with whom]$_i$ the girl talked ϵ_i	from speaking with the girl
4a	الرجل الذي انهار منزله	the+man who$_i$ collapsed house+his$_i$	the man [whose house]$_i$ ϵ_i collapsed	a man who collapsed home
4b	رجلا انهار منزله	man ω_i collapsed house+his$_i$	a man [whose house]$_i$ ϵ_i collapsed	a man of his house collapsed
4c	من انهار منزله	who$_i$ collapsed house+his$_i$	[whose house]$_i$ ϵ_i collapsed	of his house collapsed
5	ما هو منطقي	what$_i$ it$_i$ logical	what$_i$ ϵ_i is logical	what is logical

Table 1: A list of MSA sentences starting with relative clauses أعرف (translation: I know) along with their English glosses, English reference translation (RT), and the output of MT system X. Empty categories are indicated with ϵ and empty WH nodes are indicated with ω. Subscripts indicate coreference. To avoid clutter, the glosses do not explicitly indicate person, number, or gender.

antecedent plays the role of the subject, no resumptive pronoun is inserted[1]; instead, MSA inflects the verb to agree with the subject in number and gender by attaching an affix[2]. A second significant difference between the two languages is that, in MSA, relative pronouns are required for relative clauses modifying definite noun phrases but are prohibited when modifying indefinite noun phrases; in English, definitiveness neither prevents nor necessitates the inclusion of a relative pronoun. A third significant difference is that, for free relative clauses—that is, relative clauses that are not attached to an external antecedent—MSA has a different set of relative pronouns for introducing the clause[3]. A fourth challenge is that MSA has no equivalent word for the English word 'whose' and, to convey a similar meaning, employs resumptive pronouns as possessive modifiers. Examples illustrating these differences are provided in Table 1. For further background on MSA relative clauses and MSA grammar, we refer readers to books by Ryding (2005) and Badawi et al. (2004).

3 Data

In our research, we rely on the conversion of constituent into dependency structures and the training/dev/test splits of the Arabic Treebank (ATB) parts 1, 2, & 3 (Maamouri et al., 2004; Maamouri and Bies, 2004) as presented by Tratz (2013). We extract features from labeled dependency trees (rather than constituent trees) generated by Tratz's (2013) Arabic NLP system, which separates clitics, labels parts-of-speech, produces dependency parses, and identifies and labels affixes.

The original ATB dependency conversion does not mark pronouns for resumptiveness, so we modify the conversion process to obtain this information. The original ATB constituent trees mark this by labeling WHNP nodes and NP nodes with identical indices. If the NP node corresponds to a null subject and the head of the S under the SBAR is a verb, we mark the inflectional affix on the verb, which agrees with the subject in gender and number, as resumptive. These inflectional affixes are included as their own category within our analyses since their presence precludes the appearance of another resumptive pronoun within the relative clause (e.g., as a direct object).

The total number of resumptive pronouns and "resumptive" inflectional affixes in the training, dev, and test sections are presented in Table 2. In

[1] A notable exception to this rule is for equational sentences. MSA lacks an overt copula corresponding to the English word 'is' and, to convey a similar meaning, resumptive subject pronouns must be inserted in these contexts.

[2] In standard VSO and VOS constructions, the verbs inflect as singular regardless of the number of the subject.

[3] These pronouns are also employed to introduce questions.

	Training	Dev	Test
Pronouns	5775	794	796
Inflectional affixes	6161	807	845

Table 2: Number of resumptive pronouns and "resumptive" inflectional affixes by data section.

the training data, the four most likely positions[4] for the resumptive pronouns are:

i) direct object of relative clause's main verb (33.9%)
ii) object of a preposition attached to the verb (20.8%)
iii) possessive modifier of the subject of the verb (5.4%)
iv) subject pronoun in an equational sentence (4.2%).

4 Translation Error Analysis

As an exploratory exercise to gain insight into the types of errors generated by current MT engines when translating from a language that inserts resumptive pronouns (i.e., MSA) to one that doesn't (i.e., English), we worked with a native Arabic speaker to produce a list of Arabic sentences that vary in terms of definitiveness (and existence, as with free relatives) of the external antecedent, and the syntactic position of the resumptive pronoun, along with English glosses and reference translations for these sentences. This set was then processed using a popular online translation system, which we refer to as system X. The sentences, their glosses, reference translations, and automatic translations are presented in Table 1.

Although system X did not typically produce English pronouns corresponding to the resumptive pronouns in the source, most of the translations proved problematic, with many of the issues being related to reordering. Thus, while system X appears to be good at not translating resumptive pronouns, its performance on the relative clauses that contain them has ample room for improvement. Our working hypothesis is that system X's English language model is effective in discounting candidate translations that keep the resumptive pronoun.

As a second exploratory exercise, we automatically extracted all the resumptive pronoun examples in the training section of the data described in Section 3 and grouped them based upon the sequence of dependency arc labels from the resumptive pronoun up to the head of the relative clause and the first letter of the POS tag of the intervening words (e.g., 'N' for noun, 'A' for adjective). For each of the thirty most common configurations, we took ten examples (for a total of 300), ran them through system X's Arabic-English model and gave both the translation and the source text to our native Arabic expert. Our expert examined whether 1) the translation engine generated a pronoun corresponding to the source side resumptive pronoun and 2) whether the translation was correct locally within the relative clause (whether the pronoun was retained or not)[5]. The results for these two judgments are presented in Table 3.

		Corresponding Pronoun?	
		Yes	No
Correct?	Yes	17	189
	No	20	74

Table 3: Expert judgments

Our expert concluded that a corresponding English pronoun was produced in only 37 of the 300 examples (12.3%). Seventeen of these were judged correct, although in many of these cases a significant portion of the relative clause was translated incorrectly even though a small portion including the pronoun was translated properly, making judgment difficult. Our expert noted that many of the correct translations involved switching the voice of the verb in the relative clause from active to passive voice using a past participle. Of the 189 that had no corresponding pronoun and were judged correct, 46 (24.3%) involved switching to passive voice. In general, it appears that system X does a good job at not generating English pronouns corresponding to MSA resumptive pronouns, although it makes numerous mistakes with the data we presented to it.

5 System Description

Our MSA resumptive pronoun identification system processes one sentence at a time and relies upon the (averaged) structured perceptron algorithm (Collins, 2002) to rank the feasible actions. When processing a sentence containing n pronouns and affixes, a total of n iterations are performed. During each processing iteration, the system considers two actions for every unlabeled

[4] Examples of these frequent configurations are in Table 1.

[5] This latter task was challenging, but permitted, as intended, lenient judgment of the MT output.

Function Definitions:
path(x) – returns a list of dependency arcs from x up through the first 'ripcmp', 'rcmod', or 'ROOT' arc (link from affix to the core word is also treated as an arc)
rDescendants(x) – returns a list of paths (dependency arc lists) from x to each descendant already marked as resumptive
pDescendants(x) – returns a list of paths (dependency arc lists) from x to each pronoun / verbal inflectional affix, not following 'cc', 'ripcmp', or 'rcmod' arcs
hasDepArc(x,y) – returns a Boolean value indicating if an arc with label y descends from x
pathToString(x) – concatenates the labels of the arcs in a list to create a string
last(x) – returns the last element in the list x
split(x, y) – splits a string x apart wherever it contains substring y, returning these pieces
deps(x), *parent(x)* – return dependency arc(s) of which x is the {head, child}
head(x), *child(x)* – returns the {head, child} of arc x
pro(x) – if x is an affix, the word attached to it is returned, otherwise x is returned
l(x) – return the label/part-of-speech for a dependency arc, affix, or word
T(x), *t(x)*, *suffixes(x)* – return the {type ('affix' or 'pro'), written text, suffixes} for x
n(x,y) – returns the word node that is y words after *pro(x)*

Given: p – pronoun or inflectional affix

Pseudocode:

```
'0:'+T(p), '1:'+t(p), '2:'+l(p), '3:'+l(parent(p)), for(s in split(l(p),'_')) { '4:'+s }
if(T(p)='affix') { for(a in deps(pro(p))) { '5:'+l(a) }, if(T(p)='pro' or not(hasDepArc(pro(p), 'subj'))) { '6' }
for(i in {-3,-2,-1,0,+1,+2,+3,+4}) { '7:'+i+t(n(pro(p),i)), '8:'+i+l(n(pro(p),i)), '9:'+i+l(parent(n(pro(p),i))) }
'10:'+pathToString(path(p)), end := last(path(p)), resumptives := rDescendants(child(end))
if(l(end) != 'ROOT') {
    if(size(resumptives) > 0) {'11a' } else {'11b'+(size(pDescendants(child(end))) > 0)}
    for(s in split(l(head(end)), '_')) '12:'+s, for(arc in path(p)) { '13'+l(arc) }
    '14:'+t(head(end)), '15:'+l(head(end)), '16:'+l(parent(head(end)))
    '17:'+t(child(end)), '18:'+l(child(end)), '19:'+l(parent(child(end)))
    if(l(child(end)) = 'VB_PV' and size(suffixes(child(end)))=0) { '20' }
    for(suff in suffixes(head(end))) { for(s in split(l(suff), '_')) { '21:'+suff }} }
```

Figure 1: Pseudocode for feature production. Statements in bold font produce strings that are used to identify features. The feature set consists of all pairwise combinations of these strings.

personal pronoun and inflectional verbal affix[6] within a given sentence, these actions being *label-as-"resumptive"* and *label-as-"not-resumptive"*. The highest scored action is performed and the newly-labeled pronoun or affix is removed from further processing.

The system scores each action by computing the dot product between the feature vector derived for the pronoun/inflectional affix and the weight vector. The feature vectors consist entirely of Boolean values, each of which indicates the presence or absence of a particular feature. Each feature is identified by a unique string and these strings are generated using the pseudocode presented in Figure 1. (All pairwise combinations of the strings generated by the pseudocode are included as features.)

For space reasons, we omit a review of the training procedure for the structured perceptron and refer the interested reader to work by Goldberg and Elhadad (2010).

6 Experiments

We trained our system on the training data using the gold standard clitic segmentation, parse, and part-of-speech information and optimized it for overall F1 (pronouns and inflectional affixes combined) on the development data. Performance peaked on training iteration 8, and we applied the resulting model to two treatments of the test data, once using the gold standard annotation and once using the Tratz (2013) Arabic NLP system to automatically pre-process the data.

6.1 Results and Discussion

The scores for the development and test sections, both for gold and automatic annotation, are presented in Table 4.

The system performs well when given input with gold standard clitic segmentation, POS tags, and dependency parses, achieving 91.9 F1 for resumptive pronouns on the test set and 95.4 F1 for the affixes. Performance however degrades substantially when automatic pre-processing of the source is input instead. Some of this drop can be explained by the use of gold standard markup in training—more weight was likely assigned to

[6] Occasionally an imperfect verb will have both a written inflectional prefix and a written inflectional suffix. For these cases, the system only considers the prefix as there is no need to make two separate judgments.

		Pronoun			Inflectional Affix		
		P	**R**	**F1**	**P**	**R**	**F1**
Dev	Gold	92.5	92.8	92.6	96.7	96.4	96.5
	Auto	88.0	81.0	84.4	86.1	77.3	81.5
Test	Gold	92.1	91.7	91.9	95.0	95.9	95.4
	Auto	83.6	72.8	77.8	86.6	76.0	81.0

Table 4: Precision, recall, and F1 results for the "is-resumptive" label on the development and test sets for gold standard clitic separation/POS tagging/parsing and automatic preprocessing.

parse and POS tag-related features than would have if automatic pre-processing of the source had been used in training.

Having examined the classification system errors on the development data, we conclude that the main source of this drop is due to poor identification and attachment of bare relatives[7] by the Tratz (2013) NLP system. While the NLP system achieves 88.5 UAS and 86.1 LAS on the development section,[8] its performance on identifying bare relatives is comparatively low, with 70.0 precision and 60.5 recall. For the test section, the NLP system performance on bare relatives is even lower at 69.6 precision and 52.7 recall. This helps to explain why our resumptive pronoun classifier performs worse on the test data than on the development data when using automatic pre-processing but not when using gold standard markup.

7 Related Work

The computational linguistics research most relevant to ours is the work on identifying empty categories for several languages, including English, Chinese, Korean, and Hindi. Empty categories are nodes in a parse tree that do not correspond to any written morpheme; these are used to handle several linguistic phenomena, including pro-drop. Recent research demonstrates that recovery of empty categories can lead to improved translation quality for some language pairs (Chung and Gildea, 2010; Xiang et al., 2013). For more information on the recovery of empty categories, we refer the interested reader to work by Kukkadapu and Mannem (2013), Cai et al. (2011), Yang and Xue (2010), Gabbard et al. (2006), Schmid (2006), Dienes and Dubey (2003), and Johnson (2002).

8 Conclusion

In this paper, we present the challenge of translating MSA relative clauses, which often contain resumptive pronouns, into English, which relies on (inferred) empty categories instead. We examine errors made by a popular online translation service on MSA relative clauses and present an automatic system for identifying MSA resumptive pronouns.

The online translation service occasionally generates English pronouns corresponding to MSA resumptive pronouns, producing resumptive pronouns for only 37 of 300 examples that cover a variety of frequent MSA relative clause structures.

Our MSA resumptive pronoun identification system achieves high levels of precision (92.1) and recall (91.7) on resumptive pronoun identification when using gold standard markup. Performance drops significantly when using automatic pre-processing, with precision and recall falling to 83.6 and 72.8, respectively. One of the sources of the drop appears to be the weak performance of the Tratz (2013) Arabic NLP system in identifying and attaching bare relative clauses—that is, relative clauses that lack a relative pronoun.

This work is the first attempt we are aware of to automatically identify resumptive pronouns in any language, and it presents a baseline for comparison for future research efforts.

9 Future Work

Going forward, we plan to experiment with applying our resumptive pronoun identifier to enhance MT performance, likely by deleting all resumptive pronouns during alignment and, again, at translation time. Another natural next step is to train the system using automatically generated parse, part-of-speech tag, and clitic segmentation information instead of gold standard annotation to see if this produces a similar drop in performance. We also plan to investigate the use of frame information of Arabic VerbNet (Mousser, 2010) as features, and we would like to focus in greater detail on the difficulties in generating resumptive pronouns when translating from English into MSA.

References

Elsaid Badawi, Michael G. Carter, and Adrian Gully. 2004. *Modern Wrtitten Arabic: A Comprehensive Grammar*. Psychology Press.

[7] Relative clauses lacking a relative pronoun. As explained in Section 2, MSA lacks relative pronouns for relative clauses modifying indefinite noun phrases.

[8] UAS and LAS stand for unlabeled and labeled attachment scores.

Shu Cai, David Chiang, and Yoav Goldberg. 2011. Language-independent parsing with empty elements. In *ACL (Short Papers)*, pages 212–216.

Tagyoung Chung and Daniel Gildea. 2010. Effects of empty categories on machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 636–645.

Michael J. Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and experiments with Perceptron Algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

Péter Dienes and Amit Dubey. 2003. Antecedent recovery: Experiments with a trace tagger. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 33–40.

Ryan Gabbard, Mitchell Marcus, and Seth Kulick. 2006. Fully parsing the penn treebank. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 184–191.

Y. Goldberg and M. Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *HLT-NAACL 2010*.

Mark Johnson. 2002. A simple pattern-matching algorithm for recovering empty nodes and their antecedents. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 136–143.

Puneeth Kukkadapu and Prashanth Mannem. 2013. A statistical approach to prediction of empty categories in hindi dependency treebank. In *Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, page 91.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages*, pages 2–9.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan 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.

Jaouad Mousser. 2010. A Large Coverage Verb Taxonomy for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

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

Helmut Schmid. 2006. Trace prediction and recovery with unlexicalized pcfgs and slash features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 177–184.

Stephen Tratz. 2013. A cross-task flexible transition model for arabic tokenization, affix detection, affix labeling, pos tagging, and dependency parsing. In *Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*.

Bing Xiang, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the Ghost: Modeling Empty Categories for Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic*.

Yaqin Yang and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the chinese treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1382–1390.