A Hybrid Noun Phrase Translation System

Ahmed R. Nabhan  
Faculty of Computers and Information, Fayoum University, Egypt.  
E-mail: ahmed.nabhan@gmail.com

Ahmed Rafea  
Computer Science Dept., American University in Cairo, 113, Sharia Kasr El-Aini, P.O. Box 2511, 11511, Cairo, Egypt.  
E-mail: rafea@aucegypt.edu

Abstract

We present a hybrid based Noun Phrase (NP) translator that combines rule-based transfer technique with a statistical n-gram language model for selecting the best translation. Noun Phrase is the dominating construct in natural language text and targeting it for focused processing increases effectiveness of language processing systems. Manipulation of Noun Phrases is an effective subtask in Statistical Machine Translation, Multilingual Information Retrieval and Information Extraction. In this work, we make use of knowledge about Arabic language morphology regarding the translation of Verbal Nouns (Masader) and Annexation Constructs.

1. Introduction

Towards the goal of getting reliable statistical translation results, researchers focus their efforts on enhancing the way different tasks of machine translation are performed. Some researchers focus on modeling, with emphasis on innovating better models for word based, phrases based, and syntax based statistical machine translation. Other researchers focus on the development of better decoding algorithms. Another active research area is the development of reliable evaluation methods. However, little effort has been dedicated to improve the training process, despite its effect on translation quality. The Estimate Maximization (EM) algorithm has been used so long for estimating model parameters. Little effort was dedicated to innovating better training algorithms. Most algorithms of statistical translation are imported from the speech recognition area.

Noun Phrase (NP) is a very common constituent in natural languages. It has useful structures that allow us to express concepts and terms. NP is composed of a head noun that may be modified by a number of modifiers for a variety of purposes. For instance, a head noun may be followed by a second noun to specify, define, or show agent relationship with the first noun (K. C. Ryding, 2005). Modifiers enable NPs to be widely used in textual materials. In a preliminary experiment, we used the Arabic language processing tools developed by (Diab et al., 2004) to process a corpus of 1500 Arabic sentences. There were about 37,000 base phrase chunks of which base noun phrases were 22,000 phrases, representing about 60% out of the base phrases.

In this paper, we present a hybrid Noun Phrase translator that combines rule-based machine translation technique with statistical models. As the term hybrid suggests, we combine statistical-based models with rule-based ones to keep the benefit of a controlled translation process enhanced with statistical models, especially in process of lexical disambiguation. NP structures are usually easy to align with each other in parallel corpora, and in some language pairs, NP chunks are monotonic and, therefore, are easily translated. Nevertheless, some forms of NPs in Arabic can be translated accurately with insight understanding of Arabic noun term structures and the formulations of some adequate rules that handle forms as annexation and attached pronouns. The rule-based effect outstands in a set of word reordering grammar rules, lexical translation of some regular Arabic noun patterns in NP annexation structure, with the possibility of introducing spurious words like ‘of’ which does not always have Arabic equivalents, and finally the treatment of attached pronouns in Arabic nouns.

One characteristic of the proposed system is the utilization of monolingual corpora for lexical disambiguation. We built a statistical n-gram language model from the English side of Arabic-English news corpus (LDCTM2004) and used this model to rank a set of translation candidate generated by the rule-based translator. Although Arabic and English have plentiful bilingual corpora, these corpora are not always available for every genre of text (other than news) and not available for every domain (Medicine, Agriculture, Economy…).
The proposed translator needs input text to be processed for syntactic analysis and in-text annotations of syntactic information must be present in the document being processed. At the first step, words for each NP are reordered according to a set of manually developed rules. At the second step, we lookup Arabic words in the dictionary and generate a set of English phrase translation candidates, according to how many possible word translations exist for an Arabic word. At the third step, the set of translation candidates are ranked using a statistical language model and the translation candidates with the best score is output as the best translation.

We evaluate the proposed system on Arabic news text. We compare the output of this system with a set of manual reference translations for NPs that are present in text corpus. The proposed NP translator outperforms general statistical machine translation systems in terms of BLEU score and the system may be used as a subtask of a sentence translation process.

The paper is organized as follows. In section 2, we present background on some aspects of Arabic noun phrase structure that motivates our work. Section 3 presents the model and implementation of our system. Section 4 presents empirical results and Section 5 makes discussion on these results. Sections 6 concludes the paper and suggests future research needed.

2. Background

Noun phrase translation is an attractive task in machine translation systems, multilingual information retrieval and multilingual information extraction systems. Usually noun phrases represent terminology and concepts in textual material. Noun phrases constitute queries in Information retrieval systems. Information extraction systems manipulate noun phrases and question answering systems looks for key noun phrases to compose an answer. A particular text processing application may focus on the manipulation of noun phrase from many perspectives. Manipulation of noun phrases varies from chunking, extraction, and translation.

Multilingual text processing applications benefit from noun phrase translation as a unit task. Multilingual information retrieval makes use of noun phrase translation services to translate terms back and forth between language pairs. In machine translation, noun phrase translation may be handled separately as a subtask. Some statistical decoders (Koehn et al., 2007) allow the inclusion of XML tags, containing translations of noun phrases in input text, as a preprocessing step for translation.

There have been a few attempts to translate noun phrases as a sub-task of sentence translation. Koehn and Knight (2003) proposed a technique for translating noun phrases using a bilingual corpus of noun phrases that is extracted from a word-aligned and phrase-chunked corpus. The results they reported show improvements in translation quality. Despite their success, no further work was reported about NP translation as a subtask and we are not aware of other work in the literature discusses this issue.

Shaalan et al (2000) implemented English to Arabic machine translation system for translating noun phrases in titles and abstracts of scientific publications. They use a pure transfer-based approach with DCG grammar for the analysis of noun phrases.

Availability of high-precision chunking tools for Arabic language (Arabic Support Vector Machine Tools (ASVMT) (Diab et al., 2004), with a good precision score of 93% for chunking task) makes it possible for us to reliably identify noun phrases in Arabic text. This encouraged us to develop a machine translator focusing on noun phrases

3. System Description

The system is a Java application with a pipeline architecture composed of three steps. Firstly, we use a hand-crafted set of reordering patterns to put words in the right English order. The second step is to lookup Arabic words in a bilingual dictionary, generating bag of possible English translations for each Arabic word. At this step, we generate a set of translation candidates of word combinations. Finally, we use a Bi-gram language model to assign a score to every translation candidate, and then select the English translation candidate with the best score.

System description is organized as follows. Section 3.1 describes the effort of producing restructuring rules. Section 3.2 describes the NP translation system at runtime.

3.1 Developing Restructuring Rules

3.1.1 Analyzing NP Patterns

The rule-based component of the system is a hand-crafted set of word reordering rules that handle structural differences between Arabic and English. To prepare this set, we investigated NP chunks that are annotated and
extracted automatically from a development set of 1500 Arabic sentences. We extracted this development set from LDC news corpus (LDCTM2004).

The first step towards development of reordering rules is to analyze an Arabic corpus syntactically and extract NP structures. For analysis and identification of NPs in Arabic corpus, we used high precision Arabic processing tools ASVMT (Diab et al., 2004). Basically, ASVMT analyzes text documents and then produces in-text annotations of base phrase structures. Annotations include necessary information about part-of-speech and phrase boundaries. Identifying phrase chunks is a final output of pipeline process where input text is manipulated through a series of steps including tokenization, lemmatization, part of speech tagging, and finally phrase chunking. The final output is a text file of annotated Arabic sentences.

Regular expressions are then used to extract NP chunks from annotated text into a list of NPs. We produce frequency list of NP constructs and sort frequency list in a descending order. The most frequent NP structures in the list were studied and reordering rules were made. Table 1 represents frequent NP patterns together with their frequency list in a descending order. The most frequent NP structures in the list were studied and reordering rules were made. Table 1 represents frequent NP patterns together with their percentage relative to other phrase structures that exists in text.

```
<table>
<thead>
<tr>
<th>Base Phrase Pattern</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>[NP NN JJ]</td>
<td>48%</td>
</tr>
<tr>
<td>[NP NN NN]</td>
<td>13%</td>
</tr>
<tr>
<td>[NP NNS JJ]</td>
<td>8%</td>
</tr>
<tr>
<td>[NP NNP NNP]</td>
<td>8%</td>
</tr>
<tr>
<td>[NP NN JJ JJ]</td>
<td>8%</td>
</tr>
<tr>
<td>[NP DT NN]</td>
<td>3%</td>
</tr>
<tr>
<td>[NP NNP JJ]</td>
<td>2%</td>
</tr>
<tr>
<td>[NP NN CC NN]</td>
<td>1.5%</td>
</tr>
<tr>
<td>[NP NNP NNP NNP]</td>
<td>1.5%</td>
</tr>
<tr>
<td>[NP NNS JJ JJ]</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
```

Table 1: Frequency of some Base NP in Dev Set

Table 1 shows that, from a 1000 NP construct that were extracted from the development set, nearly 50% of them were simple NN JJ pattern (a singular Arabic noun followed by an adjective). For each noun phrase structure pattern, we write exactly one reordering rule. Our system has a set of 120 reordering rules. The patterns exist in Table 1 totally represents 94.5% of NP constructs found in development set. These patterns count to 10 phrase patterns. This means that 110 patterns of 120 only represent 5.5% of NP constructs in dev set.

### 3.1.2 Developing Reordering Rules

We use a quite simple reordering method for Arabic NP structural transfer into English NP structure. We create a set of deterministic reordering patterns, i.e. one rule per Arabic NP pattern. We do take some exceptions into consideration, especially when dealing with annexations and attached pronouns. That is, we reorder some NPs according to their internal structure and some times according to some lexical information found in word surface forms. For instance, we may handle a structure based on the definiteness of nouns. Phrase reordering rule for a given Arabic NP is induced from the word order in the English correspondence that exists in the English side of the parallel corpus. In table 2 we present a sample of reordering patterns used in our system.

```
| [NP NN JJ CC JJ] | 2,3,4,1 |
| [NP NN NN JJ]   | 1,3,2   |
| [NP NN JJ]      | 2,1     |
| [NP NN NNS PRP$] | 1,3,2   |
| [NP NN PRP$ JJ] | 2,3,1   |
| [NP NN CC NNS]  | 1,2,3   |
```

Table 2: A Sample Reordering Rules

A rule pattern consists of two parts, separated by an arc (→).

[Head POS POS …] → i, j, k …

The first part is a list of Tags preceded by a Head tag representing the type of the construct. The Head is the word NP. The second part is a list of integers indicating the position in English phrase where to put the word. For instance, consider the following reordering rule:

[NP NN JJ] → 2, 1

This rule states that an NP consisting of singular noun (NN) followed by and adjective (JJ) is reordered by putting the translation of (NN) in the second position in the output English phrase and putting the translation of (JJ) into the first position in the output English phrase.

The following subsections elaborate on the development of two special NP: pronouns attached to a noun, and annexation constituent.

### 3.1.3 Handling Attached Pronouns
In Arabic, it is common to attach some grammatical units such as pronouns to nouns and verbs. These grammatical units have a syntactic function in the sentence. For instance, a pronoun attached to a verb may be a subject or an object. ASVMT performs a lemmatization process in which some attached grammatical units are separated from nouns and verbs. The part-of-speech tagger assigns a special POS tag (PRP$) to attached pronouns. We handled the case of attached pronouns by making reordering rules that match their patterns in noun phrase constituents.

For instance, consider the following Arabic noun phrase.

\[\text{[NP AjtmAE/NN hm/PRP$ Almqbl/JJ]}\]

Their next meeting

This NP constituent matches the pattern

\[\text{[NP NN PRP$ JJ]} \rightarrow 2,3,1\]

This rule puts the example NP in the right order:

\[\text{[NP hm/PRP$ Almqbl/JJ AjtmAE/NN]}\]

And at the lexical transfer, the following translation takes place:

\[
\begin{align*}
\text{hm/PRP$} &= \text{their} \\
\text{Almqbl} &= \text{next} \\
\text{AjtmAE} &= \text{meeting}
\end{align*}
\]

### 3.1.4 Handling Annexation Constituents

Two nouns may be linked together in a phrase called annexation structure (iDaafa). This type is a quite common constituent in Arabic language. In annexation, the first noun has no definite article because it is in an "annexed" state and the second noun may be marked for definiteness or indefiniteness (Ryding, 2005). We translate annexation construct through three ways. One way is to keep word order unchanged and inserting 'of' between the two nouns. The second way is to keep word order unchanged and use the gerund form of the first noun. The third way is to swap translated words positions. The selection of the right way to deal with annexation takes place according to word surface form.

**Rule 1:**

If the first Arabic noun is a verbal noun (MaSdar), then that Arabic noun keeps its position in phrase and is translated as a gerund form.

Verbal nouns are used in annexation constructs to represent actions, so it is appropriate to translate them in a gerund form (Verb+ing). The noun following a verbal noun may be the doer (subject) or the object of the action (Ryding, 2005). Also, there may be two nouns following a verbal noun; one of them is the doer and the other is the object.

Table 3 shows some verbal nouns patterns and gives some examples with English translations. The letter ṡ, ʾ, ِ are the three radical letters of Arabic verbs. The letters in bold are matched as is when checking the type of verbal noun. We can easily get the verbal form of a verbal noun easily. For instance, for the verbal noun pattern (A-s- ṡ-t-f-E-A-l), the verb has the pattern (A-s-t-f-E-l) i.e. we remove the second Alef (A). Also, the pattern (m-ṣ-A-E-l-p) has a verb pattern of (f-A-E-l). Once we get the verb from the lexicon, we generate gerund form. This is achieved in the lexical transfer phase.

Consider the following example:

using weapons  AstxDAm AlAslhp

This verbal noun represents an action of the verb (use). The second noun in this NP is considered an object and the doer is implicit (unknown). The translation of this NP according to Rule 1 is to generate the verb + ing form of the verb ‘use’ since it is recognized as a verbal noun according to the rule.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-s-t-f-E-A-l</td>
<td>AstxDAm (using), AstxrAj</td>
</tr>
<tr>
<td>t-f-E-y-l</td>
<td>tødym (introducing), tznym</td>
</tr>
<tr>
<td>A-f-E-A-l</td>
<td>ArsAl (organizing), tslóm</td>
</tr>
<tr>
<td>t-f-A-E-l</td>
<td>tbAdl (delivering), t</td>
</tr>
<tr>
<td>m-f-A-E-l-p</td>
<td>m$Ahdp (watching), mAEdp</td>
</tr>
</tbody>
</table>

Table 3: Sample Verbal Noun Patterns

**Rule 2:**

If the second noun in annexation construct is definite noun (preceded by the determiner Al=the), then we keep word order unchanged and insert 'of' between the two nouns.

This rule is suitable especially when the second noun is a proper noun or a common noun.

Example:

Ministry of education  wzArp AltElim
Rule 3:
If the annexation construct is composed of two nouns in indefiniteness state, then swap the two words.

Examples:
sun glasses nzArp $ms نظارة شمس
examination committee ixtbAr AmtHAn لجنة اختبار
state policy syasp dwlp سياسة دولة

3.2 NP Translation in a Text

In the following subsections, we illustrate the process of translating a given input phrase in 3 steps.

The input of the system is a set of documents that are syntactically annotated with POS tags and phrase chunks. This is the same preprocessing procedure that we perform on development set. Next, we grasp NP structures found in the given documents. The extracted structures not only contain POS tags but also contain word surface forms. Now, we have a set of NP structures ready for translation.

Step 1: Matching input NP with Reordering rules

Giving an extracted noun phrase with surface word form plus POS tags, these input phrases are matched against reordering patterns. That is, the POS tags are matched against reordering rules. When a phrase has a match in the table, word/part-of-speech pairs are reordered according to that pattern. If no match is found, we adopt monotonic translation and the phrase word order is left intact.

Step 2: Lexical Transfer

In lexical transfer, an Arabic word is looked up in an Arabic-English dictionary, and a list of possible translations is generated. We use the dictionary available with Buckwalter Arabic morphological Analyzer (Buckwalter, 2004) to produce English lexical equivalents for each Arabic word. In this step, we do not do any lexical disambiguation. Annexation constituents of verbal nouns are lexically translated, according to sample patterns in Table 3, into Verb+ing form to express actions.

Considering the case of translating verbal nouns, we reduce the verbal noun form to the verb form using the lexical structure of the verbal nouns. In Arabic Language, there are strict rules for this mapping. Table 4 lists some known verbal noun forms and corresponding verb forms.

<table>
<thead>
<tr>
<th>Verbal Noun Pattern</th>
<th>Verb Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-s-t-f-E-A-l</td>
<td>A-s-t-f-E-l</td>
</tr>
<tr>
<td>t-f-E-y-l</td>
<td>f-E-l</td>
</tr>
<tr>
<td>A-f-E-A-l</td>
<td>A-f-E-l</td>
</tr>
</tbody>
</table>

Table 4: Conversion from Verbal Nouns to Verbs

For example, to get the translation of verbal noun (Astxdam 'using'), we match the surface form of the word against the patterns in Table 4. In this case, it matches A-s-t-f-E-A-l and the corresponding verb is A-s-t-f-E-l. Hence, we remove extra letters from the verbal nouns and get the verb form (Astxdm 'to use'). After that, we look this verb in the dictionary and get the translation and produce verb + ing form.

After producing word translations, we generate English phrase translation candidates. The number of candidate translation increases according to the number of words in Arabic phrase and the number of English translations per Arabic word. If we have a two word Arabic phrase, and the first one has m English translation and the second one has n English translation, then we have m x n translation candidates.

If there is no direct match in the lexical database, then we apply naïve stemming to the unknown word and hit the lexicon again for a match. Naïve stemming is performed by trying to remove some prefixes and suffixes presented in the Table 5.

<table>
<thead>
<tr>
<th>Affix</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>Determiner (the)</td>
</tr>
<tr>
<td>Wn</td>
<td>Plural form</td>
</tr>
<tr>
<td>Yn</td>
<td>Plural / dual form</td>
</tr>
<tr>
<td>An</td>
<td>Dual form</td>
</tr>
<tr>
<td>P</td>
<td>Feminine form</td>
</tr>
</tbody>
</table>

Table 5: Affixes that are attached to nouns

At the end of this phase, we got a bag of English translation candidates for the Arabic noun phrase being processed. What we will do next is to find some way for scoring this set of translations for quality, and select the best candidate (using language model) as output translation.
Step 3: Scoring Translations

A score is assigned to each translation candidate to represent the quality of that candidate as being adequate translation of the source Arabic NP. The score is calculated using a standard statistical n-gram language model (LM). Statistical LMs have been integrated with statistical machine translation models from the early days of statistical machine translation (SMT) (Brown et. al, 1993). Even before SMT, statistical LMs were used in Speech Recognition and Optical Character Recognition.

Following the standard practice of incorporating language models into translation process, we use a bi-gram language model to compute a score for each English candidate translation using a smoothed bi-gram language model. The scoring function normally assigns low score to odd candidate translations and high score to adequate translations.

A language model score for an English translation candidate is calculated using chain rule:

$$h_2(e^n) = \prod_{i=1}^{n} p(e_i | e_{i-1})$$

4. Experiments and Results

Experimental Data

For language model training, we used the English side of LDCTM2004 parallel corpus. We use the Arabic lexicon available with Buckwalter morphological analyzer for lexical transfer (Buckwalter 2004). For test data, we manually translated a set of 1200 Arabic phrase to create a gold standard data for evaluation. The corpus was collected from Al-Ahram newspaper online site1. Test corpus statistics are presented in Table 6:

<table>
<thead>
<tr>
<th>No. of Sentences</th>
<th>1200 Snt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sent. Length</td>
<td>35 word</td>
</tr>
<tr>
<td>Word Count</td>
<td>47664</td>
</tr>
<tr>
<td>Total Base Phrases</td>
<td>15484</td>
</tr>
<tr>
<td>No. NP Phrases</td>
<td>6559</td>
</tr>
</tbody>
</table>

Table 6: Test Corpus Statistics

Language Model Training

We use the SRI language modeling toolkit for training a Statistical N-Gram language model, with Good-Turing smoothing method (Stolcke, 1999). The toolkit is run with default parameter settings.

System Evaluation

We use BLEU scoring metrics for evaluating system performance on test corpus. We report n-gram BLEU score for bi-gram, tri-gram and quad-gram. The measured BLEU score for our system and for Microsoft Live Translation Service2 are presented in the Table 7.

<table>
<thead>
<tr>
<th></th>
<th>2-gram</th>
<th>3-gram</th>
<th>4-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed System</td>
<td>0.4864</td>
<td>0.5135</td>
<td>0.3273</td>
</tr>
<tr>
<td>Microsoft Live Translation</td>
<td>0.4219</td>
<td>0.3035</td>
<td>0.1923</td>
</tr>
</tbody>
</table>

Table 7: BLEU scores

In addition to the overall performance of the system measured in BLEU scores of quad-grams, the BLEU scoring utility reports individual n-gram scores for unigrams, bi-gram, tri-gram … etc. These individual scores help researchers to find out interesting facts. For instance, uni-gram BLEU score can be used as a measure of out-of-vocabulary rate and a measure of performance of lexical transfer. Higher order n-grams reports effectiveness of word/syntactic reordering patterns.

According to the figures in Table 6, being a specialized system for grasping NP constituents on text and translating them, the proposed specialized NP translator performs well, especially on bi-grams and tri-grams. The significance of performance of the proposed system is clear in tri-grams and quad-grams. This performance is due to the use of specialized rules that handle special characteristics of Arabic language (Verbal nouns, attached pronouns, and annexation forms).

5. Discussion

We analyzed performance based on the structure of Arabic NP patterns and the rate of success in translation. Table 7 presents most frequent patterns and percentage of success in translation task. The success rate is based on BLEU score.

1 http://www.ahram.org.eg
2 http://www.microsofttranslator.com/Default.aspx
Table 8 shows that the most frequent phrase structure (NN JJ) can be translated with a success rate of nearly 55%. A comparable success rate (52%) is achieved with (NN JJ JJ) structure. We can say that the adjective noun phrase is relatively easy to translate.

Various forms of attached pronouns compose nearly 15% (648/4414) of phrase structures that were annotated in test corpus. A good success rate, 57%, (388/648) was achieved, despite we hoped for a much higher rate. This is due to the lexical ambiguity that current language model sometimes can not resolve.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Occ.</th>
<th>Translations with score = 1.0 BLEU</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>[NP NN JJ]</td>
<td>2004</td>
<td>1106</td>
<td>55%</td>
</tr>
<tr>
<td>[NP NN PRP$]</td>
<td>574</td>
<td>318</td>
<td>55%</td>
</tr>
<tr>
<td>[NP NN NN]</td>
<td>430</td>
<td>205</td>
<td>47%</td>
</tr>
<tr>
<td>[NP NNS JJ]</td>
<td>346</td>
<td>114</td>
<td>32%</td>
</tr>
<tr>
<td>[NP NNP NNP]</td>
<td>280</td>
<td>60</td>
<td>21%</td>
</tr>
<tr>
<td>[NP DT NN]</td>
<td>192</td>
<td>120</td>
<td>62%</td>
</tr>
<tr>
<td>[NP JJ CC JJ]</td>
<td>130</td>
<td>88</td>
<td>67%</td>
</tr>
<tr>
<td>[NP NN CC NN]</td>
<td>120</td>
<td>56</td>
<td>46%</td>
</tr>
<tr>
<td>[NP NN JJ JJ]</td>
<td>228</td>
<td>120</td>
<td>52%</td>
</tr>
<tr>
<td>[NP NN PRP$ JJ]</td>
<td>110</td>
<td>70</td>
<td>63%</td>
</tr>
</tbody>
</table>

Table 8: Success rate of the Proposed System for most frequent structures

There is only one noun in the attached pronoun phrase structure, and the system was not able to give high score to the best translation. In case where there are a noun and an adjective (NN PRP$ JJ), the system achieves higher success rate (63%). We due this success rate to length of the phrase which gives the n-gram language model good chance to select good words, and therefore it is able to disambiguate word senses. Also, the proper selection between the feminine pronouns (its and her) is confusing to the system.

We manually inspected attached pronouns structures and we were able to distinguish different roles of attached pronouns which led to the low success rate. In the formulation of transfer rules, we assumed that attached pronouns represent possessive relation with the noun they were attached to. For instance:

- His supporters: AnSAr + h  
- Its role: dwr + hA  
- Their arrival: wSwl + hm

When attached pronouns are attached to verbal nouns (MaSader), they play different roles and the system can not disambiguate based on the information supplied from the syntactic chunker. Sometime they represent the doer of the action and sometime they refer to the object. For instance:

AEtA'h $\rightarrow$ AEtA' + h  
AstqbAlh $\rightarrow$ AstqbAl + h

In the example above, (AEtA') means (providing) while (AstqbAl) means (meeting). In this example, the attached pronoun represents an object, not a possessive pronoun. In such a case, the system translates that first phrase as (His providing) and the second phrase as (His meeting) which are not correct. The correct translations are (Providing him) and (Meeting him).

A low rate of successful translations of Proper nouns is clear (37%). This is due to the limited support of the lexicon to Proper nouns. Translation systems may have a specialized component to deal with Proper nouns.

Annexation structures nearly represent 15% of annotated text. The phrase structure of the form (NN NN) achieved 47% success rate. This is a fair success rate and is due to the successful manipulation of this structure according to the three rules we had set in section 3.

6. Conclusion

We conclude that translation quality of basic and frequent noun phrase structures is improved via simple hand-crafted reordering patterns in addition to standard language model at the target side. Lexical disambiguation for short phrases using language models is found to be feasible when used with specialized translation systems. Language models are reliable enough to produce good translations with the aid of insight analysis at syntactic and lexical level. In the future, we will experiment with different genre of text other than news and with different monolingual corpora sizes.
Acknowledgement
This work is supported by the American University in Cairo, Egypt.

References

Andreas Stolcke. SRILM---the SRI language modeling toolkit.


Franz J. Och. GIZA++ Release notes, 2001


