Multi-level Example-Based Arabic-to-English Translation

Kfir Bar, Nachum Dershowitz
School of Computer Science, Tel Aviv University, Israel
kfirbar@post.tau.ac.il nachumd@post.tau.ac.il

This non-structural example-based machine translation system translates sentences from **Arabic** to **English**, using a parallel corpus aligned at the sentence level. Each input sentence is fragmented into phrases. Phrases are matched to example patterns, using multiple levels of morphological information.

1 Parallel Data

1. Translation examples were extracted from the given parallel unvocalized Arabic-English corpus.
2. Examples were morphologically analyzed using the Buckwalter (2002) analyzer (version 1.0), and then part-of-speech tagged using AMIRA (Diab et al., 2004).

Word alignment is done using GIZA++ augmented with one-to-one matches using a dictionary, created based on Buckwalter glossaries, expanded with WordNet synonyms. The Arabic version of the corpus was indexed on the word, stem and lemma levels.

2 Matching

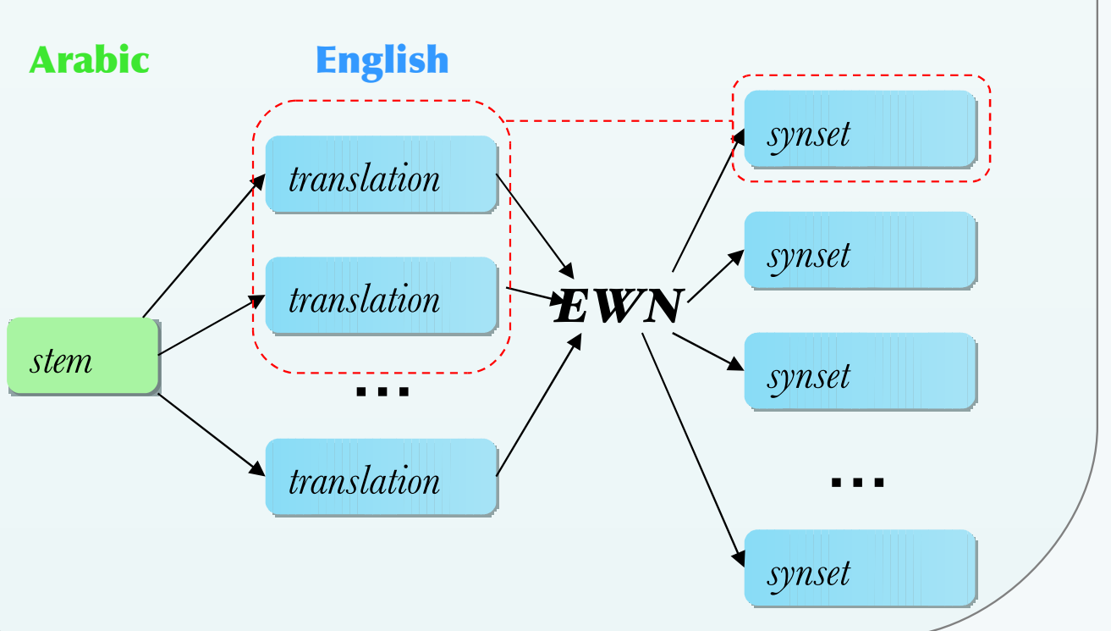
Given an Arabic sentence:

1. Search corpus for input fragments.
2. Match word-by-word at different levels, including exact match, **synonym**, stem, lemma, morphological features, proper noun. Calculate score based on all levels.
3. Matches can be found only for a combination of one or more base-phrases.

Thesaurus Creation

Every noun stem in the Buckwalter list was compared to all other stems:

- We ask the English WordNet for all (noun) **synsets** of every English translation of a stem.
- A **synset** containing two or more Buckwalter translations is a possible sense for the stem. We also considered the *hypernym* relation.

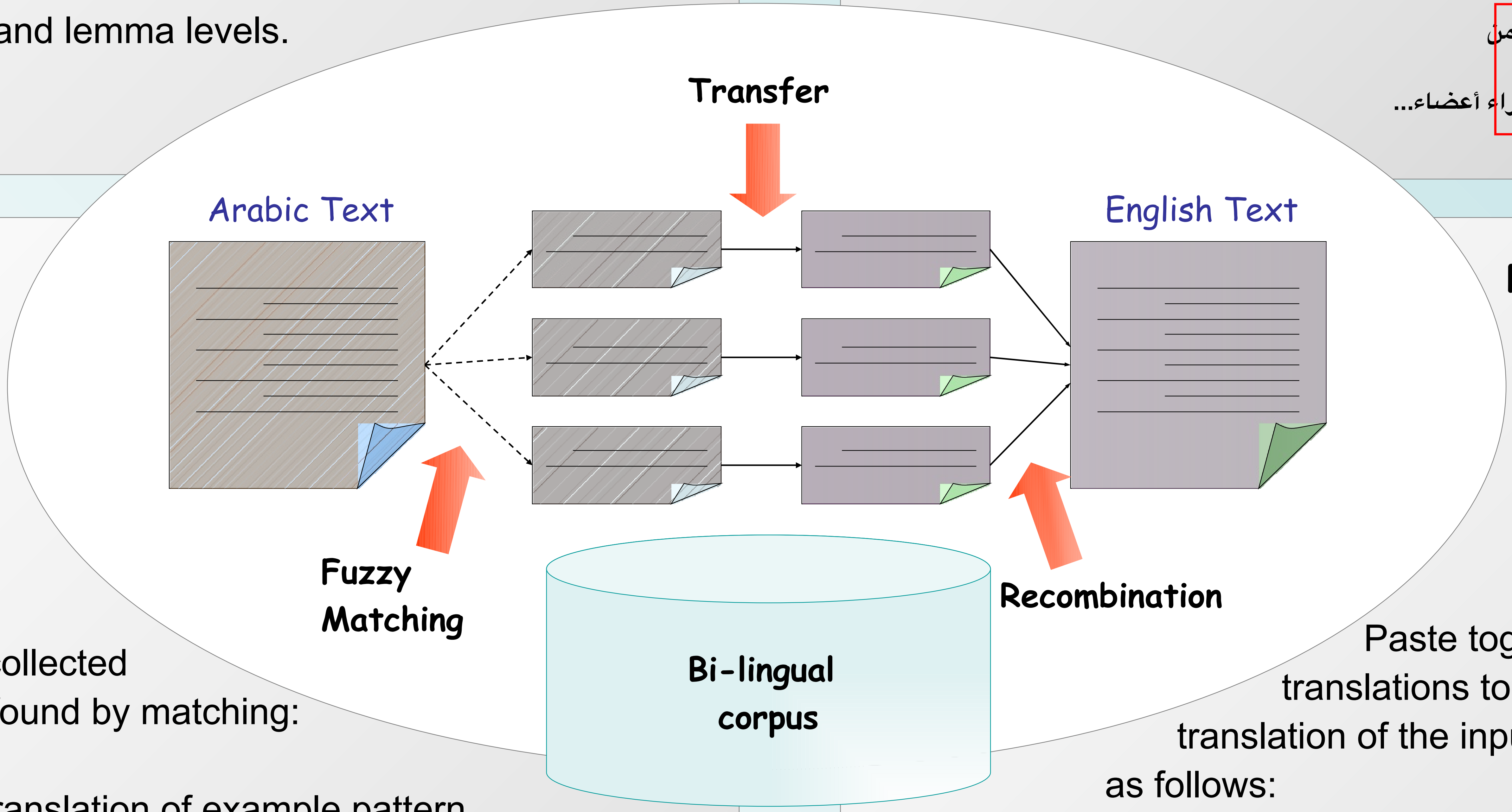


MF Match Exact Match

مذكرة من رئيس مجلس الأمن

... ويعين مجلس الوزراء أعضاء...

3 Transfer



Given the collected fragments found by matching:

1. Extract translation of example pattern from the English version.
2. Modify extracted translation to correspond to attributes of the source pattern.

Recombination 4

Paste together extracted translations to form a complete translation of the input sentence, as follows:

1. Find recombination that best covers the entire input sentence, using dynamic programming.
2. Smooth recombined translation to make it fully grammatical (in progress).

3.1

Extract shortest English fragment that is composed of maximum word translations.

Example

Arabic الخدمات الاستشارية والتعاون التقني في ميدان حقوق الإنسان
English: Advisory services and technical cooperation in the field of human rights

3.2

1. Replace translation of words matched per morph features with translation of input fragment using lexicon.
2. Remove unnecessary middle words using English shallow parser.

Example

Input fragment: موضوع الامن

Extracted translation:
the subject of regional security

After shallow parsing:
[the/DT subject/NN] of/IN [regional/JJ security/NN]

e subject of security

Results

Level / Set	DEV-1	DEV-2	DEV-3	DEV-6	DEV-7	Test-set 10	Test-set 09
Level 1	0.3672	0.3333	0.3267	0.2921	0.2800	Not Submitted	
Levels 1 – 2	0.3672	0.3333	0.3267	0.2921	0.2800		
Levels 1 – 3	0.3672	0.3334	0.3273	0.2924	0.2799		
Levels 1 – 4	0.3676	0.3333	0.3273	0.2924	0.2799		
Levels 1 – 5	0.3656	0.3333	0.3279	0.2935	0.2845	0.2321	0.2927
No synonym	0.3656	0.3332	0.3267	0.2910	0.2800	Not Submitted	

BTEC Arabic-English

Conclusions

- An initial implementation for participating in the **IWSLT 10 evaluation campaign** is presented.
- Matching words on multiple levels improves matching results and corpus exploitation.
- Using synonyms in the matching step slightly improves final results.
- Future: broadening the level of similarity using paraphrases to find new matches for the input text.