



Using Synonyms for Arabic-to-English Example-Based Translation

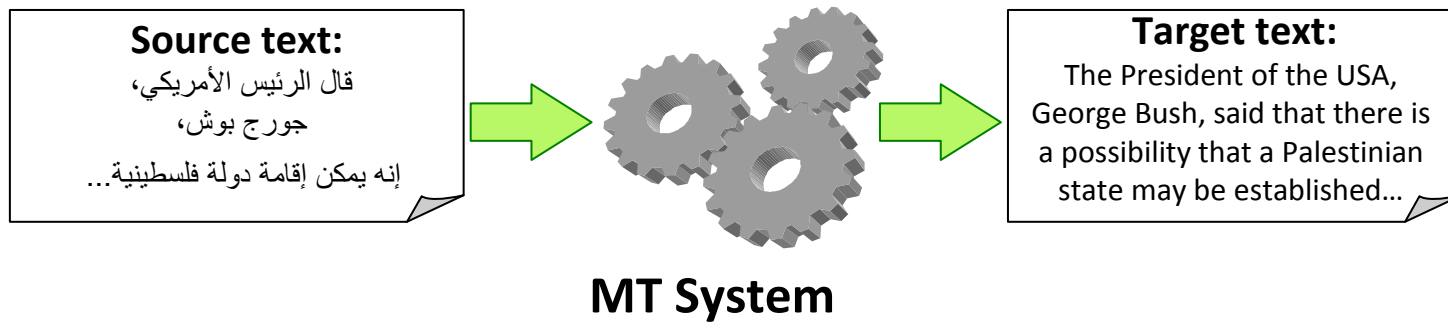
**Kfir Bar
Nachum Dershowitz**

Tel Aviv University

ISCOL 2010

Machine Translation (MT)

Use of computers to translate from one language to another



Research Paradigms



1 Rule-based

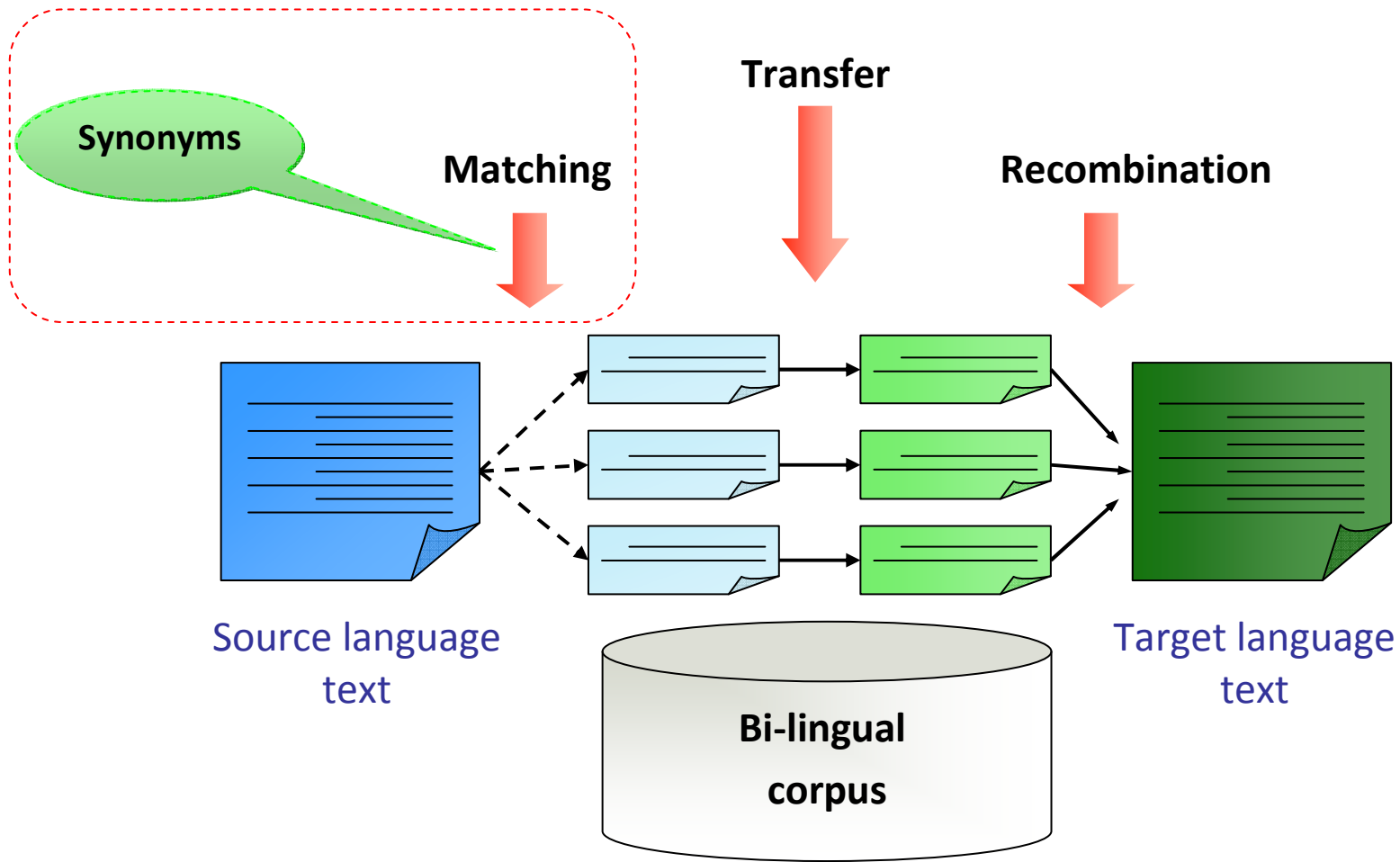
2 Corpus-based

└─┬─→ Example-based (Memory-based)

Existing EBMT Systems

- 1 M. Nagao (1984)
- 2 S. Sato & M. Nagao (1990)
- 3 E. Sumita & H. Iida (1992)
- 4 S. Nirenburg & R. D. Brown (1994)
- 5 Y. Muyun, Z. Tienjun, L. Haijie, S. Xiaosheng,
& J. Hongfei (2003)
- 6 E. Aramaki & S. Kurohashi (2003)

EBMT – Example Based Machine Translation

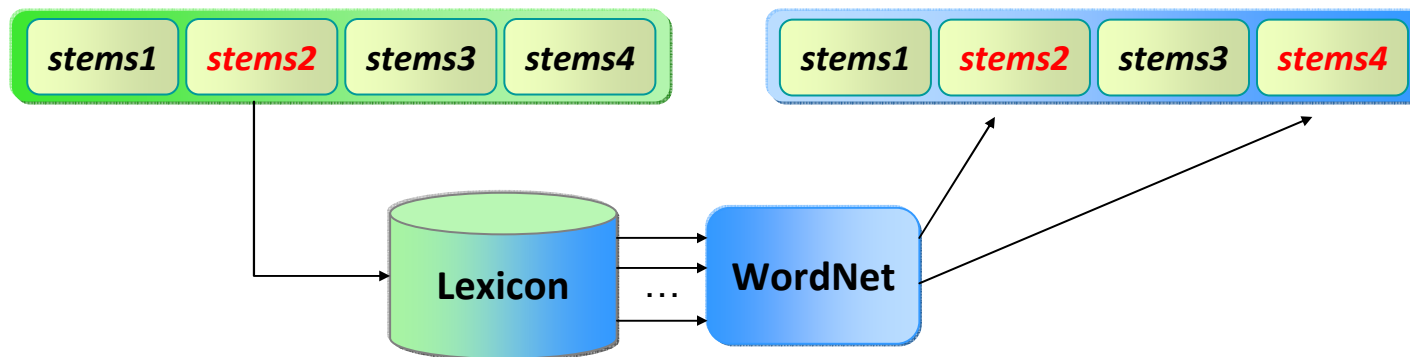


Our EBMT System

- Translates Arabic sentences into English.
- Non-structured: translation examples are stored with only some morph-syntactic information.
- Uses several parallel corpora provided by LDC.
- So far, only matching and transfer. Real recombination left for future work.

Corpus

- Uses sentence-aligned parallel corpora (by LDC).
- Translation examples were morphologically analyzed using the Buckwalter morphological analyzer, and then part-of-speech tagged using AMIRA (Diab et al., 2004).
- Creating *alignment-table* for each translation example:



Other general syntactic rules are applied for creating one-to-many alignment entries.

Matching

- Corpus is searched for input fragments.
- Matching is word-by-word at several levels.
Total score is calculated by combining level scores.



Exact match

Synonym match

Stem match

Lemma match

Morphological-feature match

- Fragment score is created from word scores.

Matching



Example:

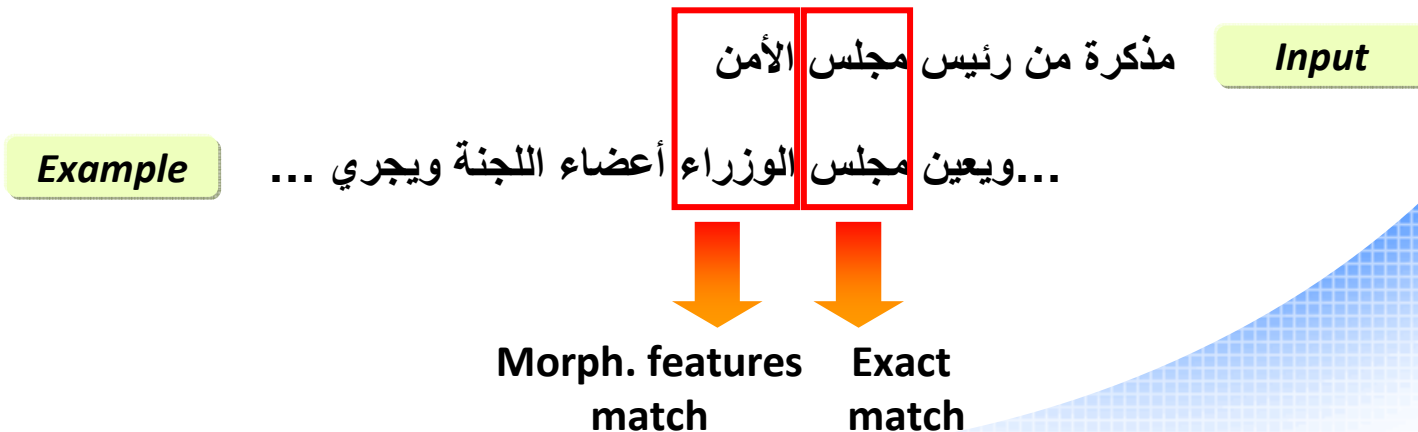
Input sentence:

مذكرة من رئيس مجلس الأمن

(A memorandum by the president of the Security Council)

Corpus example:

... ويعين مجلس الوزراء أعضاء اللجنة ويجري ...



Thesaurus Extraction

- Arabic WordNet is still under development...
- There are several works on automatic extraction of synonyms and semantically similar expressions:

Translations as Semantic Mirrors: From Parallel Corpus to WordNet, **Dyvik Helge. 2004**

Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity, **Lonneke van der Plas and Jörg Tiedemann. 2006**

Extracting Paraphrases from a Parallel Corpus, **Regina Barzilay and Kathleen R. McKeown..**

- Our current attempt uses Buckwalter lexicon and WordNet for finding **Arabic noun synonyms.**

Thesaurus Extraction

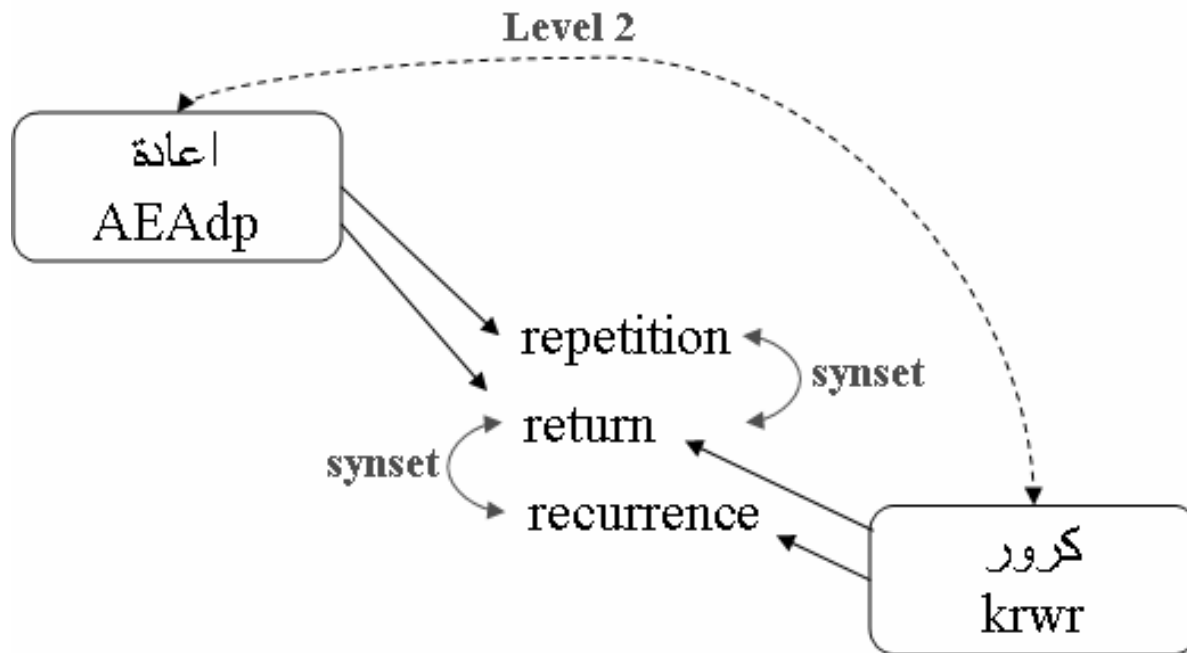
Every noun stem in the Buckwalter list was compared to all other stems

- We ask EWN 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.
- We define six levels of synonymy between **stems**:
 - 1** 2 translations in common
 - 2** 1 or more senses in common
 - 3** Same unique translation
 - 4** 1 translation each and they're synonyms
 - 5** 1 common translation
 - 6** Synonymous translations

Thesaurus Extraction



Example:



Matching

- Since words in the input sentence / corpus are not given with their senses it is difficult to match on synonyms.



We classify each input sentence by topic, as well as all the corpus translation examples. We consider synonyms only if the **topic-sets of both parts intersect**.

Classification

- We trained a simple classifier on English Reuters corpus.
- We used SVM on stems, removing stop words.
Accuracy: 94% for Reuters test-set (1219 documents).
- Used classifier on English half of all translation examples in our corpus.
- The Arabic part of those examples was used as a training-set for another classifier for the same topic list for Arabic (stems, ignoring stop words).

Results

	Small Corpus 29,992 translation examples				Large Corpus 58,115 translation examples			
	w/ classification		w/o classification		w/ classification		w/o classification	
	BLEU	MTOR	BLEU	MTOR	BLEU	MTOR	BLEU	MTOR
Level 1	0.0858	0.4410	0.0870	0.4450	0.0978	0.4560	0.0997	0.4606
Levels 1 – 2	0.0858	0.4410	0.0870	0.4450	0.0978	0.4560	0.0997	0.4606
Levels 1 – 3	0.0861	0.4413	0.0872	0.4452	0.0978	0.4560	0.0997	0.4607
Levels 1 – 4	0.0893 (+4.4%)	0.4465	0.0862	0.4395	0.0978	0.4561	0.0997	0.4608
Levels 1 – 6	0.0862	0.4414	0.0873	0.4453	0.0982	0.4564	0.1001	0.4611
No synonym	0.0853	0.4409	0.0877	0.4460	0.0977	0.4558	0.1001	0.4608

Testing on 586 sentences (MT-EVAL 09)

Conclusions and Future Work

- Synonyms benefit from being matched carefully by considering the context in which they appear.
- Using synonyms on a large corpus did not result in an improvement of the final results, as it did for a smaller corpus.
- Improving alignment and smoothing out the final English translation is under development.
- **Beginning to investigate the possibility of matching based on semantically-similar phrases (paraphrases).**



Thank you

Thesaurus Extraction

The resultant thesaurus contains:

22,621 nouns

- 1** 20,512 relations
- 2** 1,479 relations
- 3** 17,166 relations
- 4** 38,754 relations
- 5-6** 137,240 relations