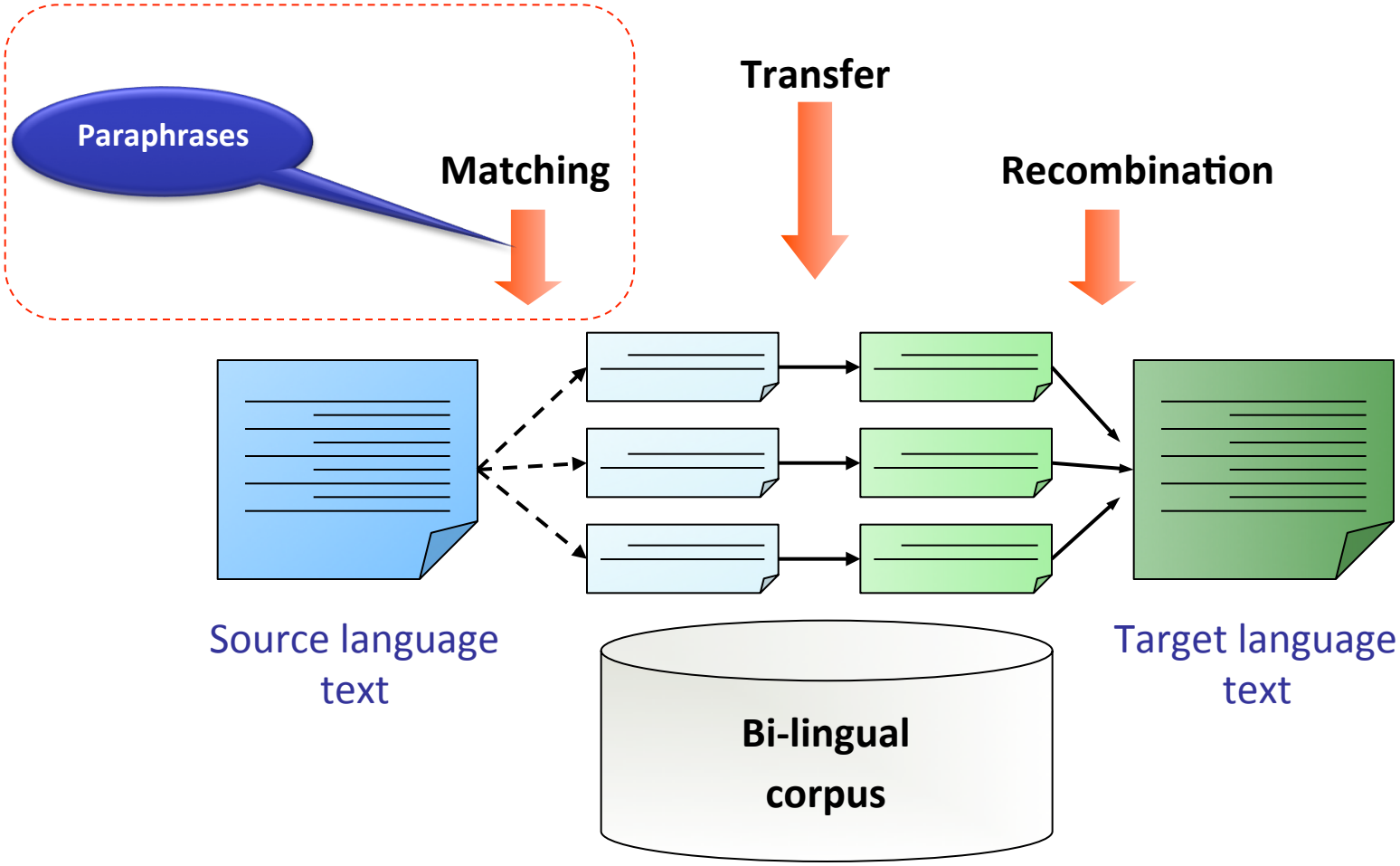
Using Verb Paraphrases in Arabic-to-English Example-Based Translation

Kfir Bar
Nachum Dershowitz

Tel Aviv University

MTML 2011

EBMT – Example Based Machine Translation



Our EBMT System

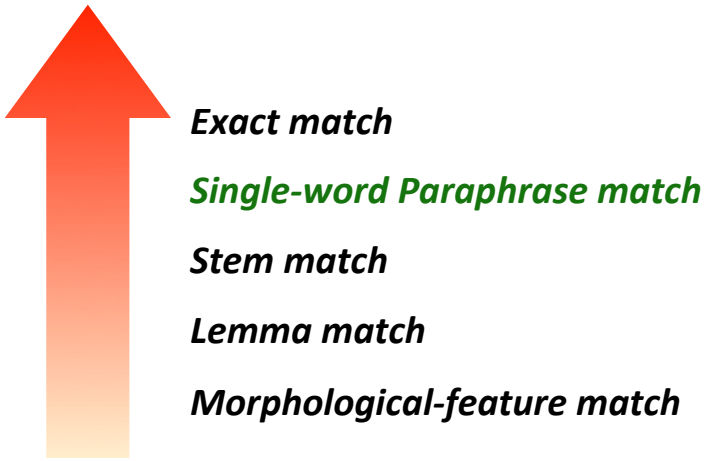
- Non-structured: translation examples are stored with only some morph-syntactic information
- Uses a parallel corpus provided by LDC
- So far, only matching and transfer. Real recombination left for future work

Corpus

- Uses sentence-aligned parallel corpus
- Translation examples were
  - morphologically analyzed using Buckwalter
  - part-of-speech tagged using AMIRA
- Word alignment by GIZA++

Matching

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



- Fragment score is created from word scores

Matching

Example

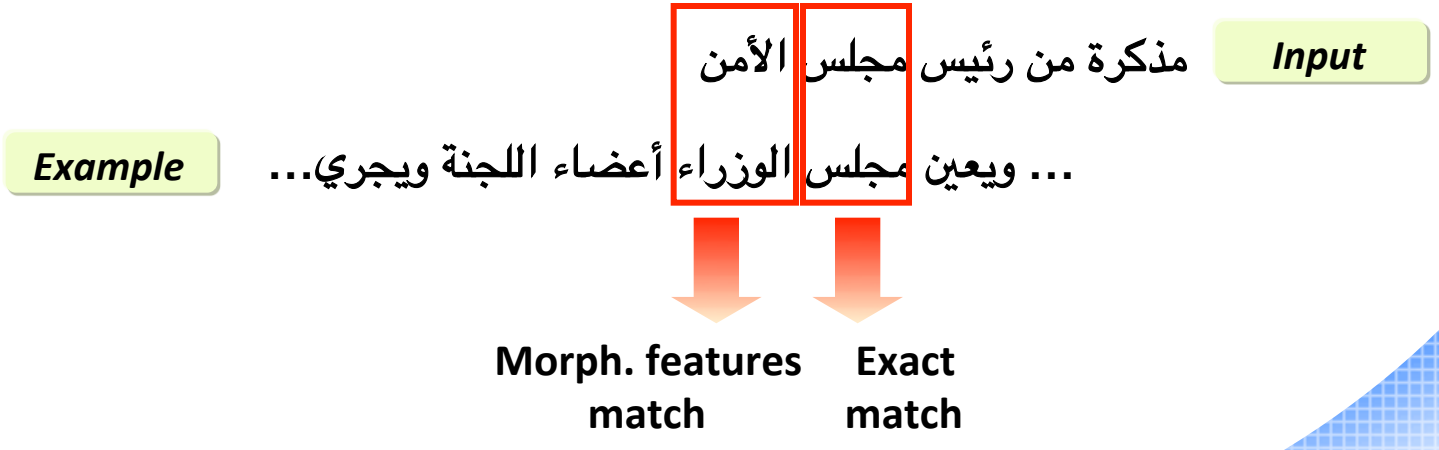
Input sentence:

مذكرة من رئيس مجلس الأمن

(A memorandum by the president of the Security Council)

Corpus example:

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



Paraphrase Extraction

- Today: single-word verb equivalents
- Inspired by:

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

- Multiple **English translations of same source**
  - Extracted multi-word paraphrases
  - Co-trained two classifiers –
    one for best paraphrase contexts
    one for best paraphrase patterns

Paraphrase Extraction

Verb Paraphrases

- Two verbs expressing the same meaning
- In at least one context the two are understood in the same way

wqd ***nql*** *AlA*AEp h*h AlAgAny*

وقد نقل الاذاعة هذه الاغاني

"convey"

wqd ***bvt*** *AlA*AEp h*h AlAgAny*

وقد بثت الاذاعة هذه الاغاني

"broadcast"

The radio broadcast these songs

Paraphrase Extraction

Context Representation

Sentence 1

mktb Alsnywrp wdywAn Owlmrt ***ynfyAn*** *xbrA En lqA' fy $rm Al$yx.*

Siniora's office and Olmert's administration **deny** a story about a meeting in Sharm al-Sheikh.

Sentence 2

mktb Alsnywrp ***ynfy*** *xbrA En lqA}h msWwlyn IsrA}ylyyn.*

Siniora's office **denies** a story about a meeting with Israeli officials.

Context

Verb: *nfy* (*ynfyAn*, *ynfy*)
Context:
Left-1 (NN, NNP) Right-1 (NN_1, IN_2)
Left-2 (NN, NNP) Right-2 (NN_1, IN_2)

Paraphrase Extraction

Arabic Verbs

root + **inflectional class (9)** = *$aAraka*

$.r.k *xaAxaxa*

Inflections

- **Gender** (*m*, *f*)
- **Person** (*1*, *2*, *3*)
- **Number** (*singular*, *dual*, *plural*)

Verb

- **Perfect** — stem: *$aArak* — lemma: *$aArak_1*
- **Imperfect** — stem: *$aArik*

Paraphrase Extraction

Arabic Verbs

- Arabic has some features that may affect paraphrasing:
  - Using different particles indicating the object may change the meaning of the verb

    qDY = to judge

    qDY ElY = to kill
  - In some cases, the verb can be replaced with its verbal noun

    zAr = to visit = *qAm* ***bzyArp***

    wqE= signed = *tm* ***twqyE***

Paraphrase Extraction

Arabic Verbs

- Sometime one cannot identify the voice, since diacritics are omitted

  kutiba (written) and ***kataba*** (to write)

- Class 7 is the passive version of class 1 and class 5 is the passive version of class 2

  qutiEa (cut - passive) = ***linqataE***

Paraphrase Extraction

Related Works

- Using parallel corpus:

  Improved statistical machine translation using paraphrases,
  Chris Callison-Burch, Philipp Koehn and Miles Osborne

- Using large monolingual corpus:

  Discovery of Inference Rules for Question Answering,
  Dekang Lin and Patrick Pantel

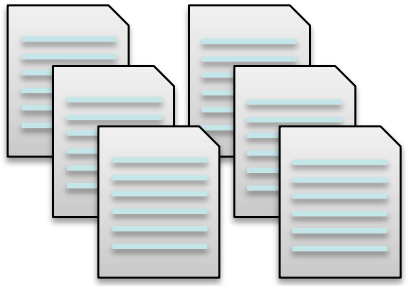
- Using corpus of comparable documents:

  Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment,
  Regina Barzilay and Lillian Lee

Paraphrase Extraction

Corpus of Comparable Documents

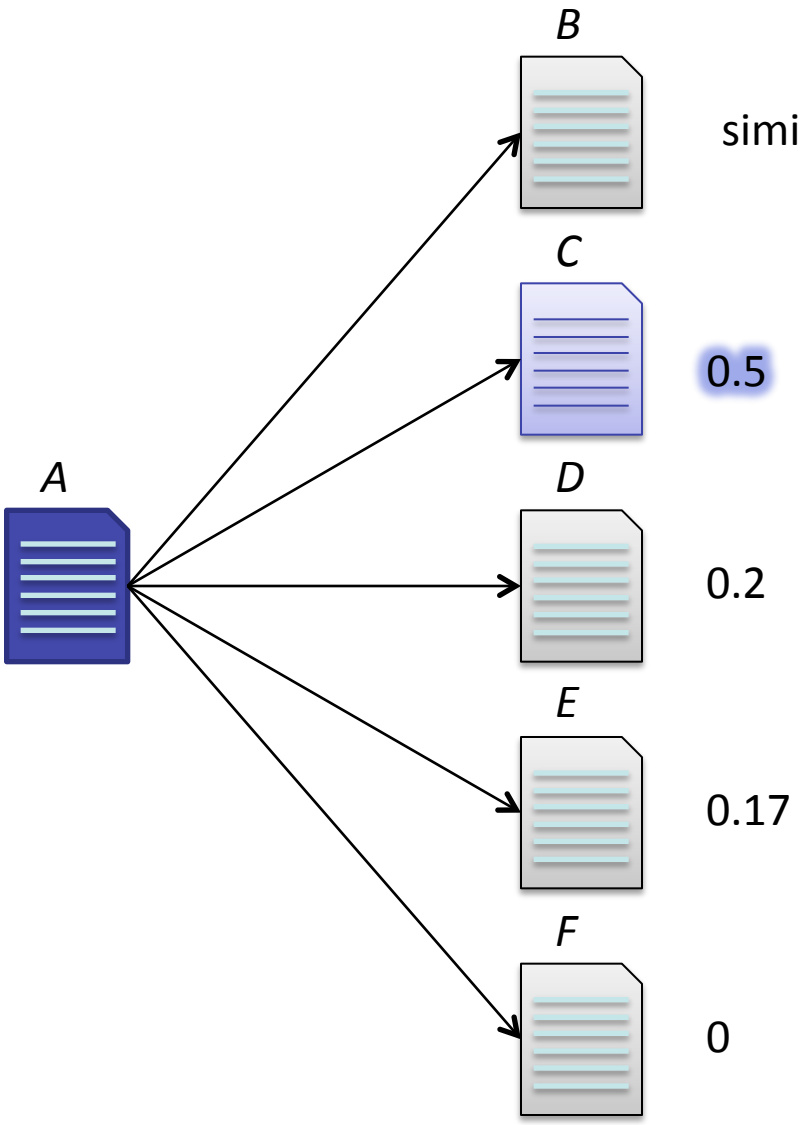
- First attempt in Arabic – using comparable corpora



- Comparable corpora extracted from the Arabic Gigaword. We automatically paired documents that were published on the same day and share title word lemmas

Paraphrase Extraction

Corpus of Comparable Documents



similar(*A*,*B*) / length(title(*A*)) = 0.2

0.5

0.2

0.17

0

Paraphrase Extraction

Context Classifier

- Extracting contexts of **positive examples** and **negative examples**, and learning the best contexts for both classes

Step 1

- We consider every verb-pair as a candidate
- We choose all candidates whose verbs share a lemma, as **positive examples**
- **Negative examples** are candidates with *unrelated* verbs
- The rest are potential candidates

Paraphrase Extraction

Context Classifier

Step 1
Cont'

We created a list of *related* verbs:

Buckwalter's **stem list** comes with English glosses

- We relate verbs that share either similar or synonymous English translations
- Synonyms were found using the English WordNet

Paraphrase Extraction

Context Classifier

Step 2

Identifying the best contexts using the **strength** and **frequency** of each context

$$strength(C) = X/N$$

X is the number of times the context ***C*** appears in a positive / negative example

N is the total number of occurrences of ***C***

Paraphrase Extraction

Context Classifier

Step 2 *Cont'*

For each class (negative / positive) we choose the ***k* most frequent** contexts with *strength* higher then a predefined threshold

k=20, *threshold*=0.95

Step 3

We consider all occurrences of the best contexts among all potential candidates

***Paraphrases* are those surrounded by a positive context but not surrounded by a negative one**

Paraphrase Extraction

Preliminary Results

We used 5500 document pairs, extracted from the Arabic Gigaword. Total number of words is ~3M.

In this experiment we use a **window of size 2**

NN NNP NN IN

wdywAn Owlmrt ***ynfyAn*** *xbrA En*

NN NNP NN IN

mktb Alsnywrp ***ynfy*** *xbrA En*

- L NN, NNP ←→ NN, NNP
- L NNP ←→ NNP
- L NNP ←→ NNP
- L NN, NNP ←→ NN, NNP
- L NN ←→ NN, NNP

- R NN_0, IN_1 ←→ NN_0, IN_1
- R NN_0, IN_1 ←→ NN_0, IN_1
- R NN_0 ←→ NN_0
- R NN_0 ←→ NN_0
- R NN_0 ←→ NN_0

…

Paraphrase Extraction

Preliminary Results

- 2 experts evaluated the results. For each candidate pair, every expert was requested to decide:

  correct – instances can be exchanged in some contexts

  incorrect – instances are not exchangeable

# of Unique Candidates	# of Unique Paraphrases	Expert 1: # of correct Paraphrases	Expert 2: # of correct Paraphrases
15,101	139	120 (86% precision)	103 (74% precision)

Paraphrase Extraction

Examples

Paraphrases

AEtql ⟺ *Owqf*

bv~ ⟺ *nq~l*

lstqbl ⟺ *lltqY*

Best Positive Contexts

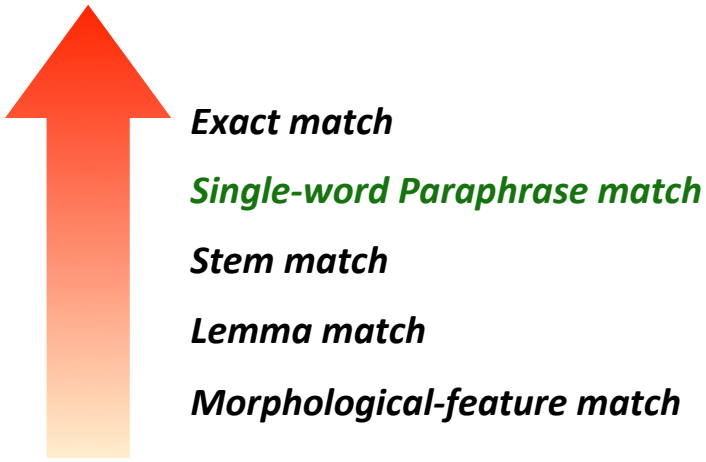
L NN_0 ←→ NN_0

R IN, NN ←→ IN

L NN, WP_0 ←→ WP_1

R NN_0←→ NN_0

Paraphrases in Translation



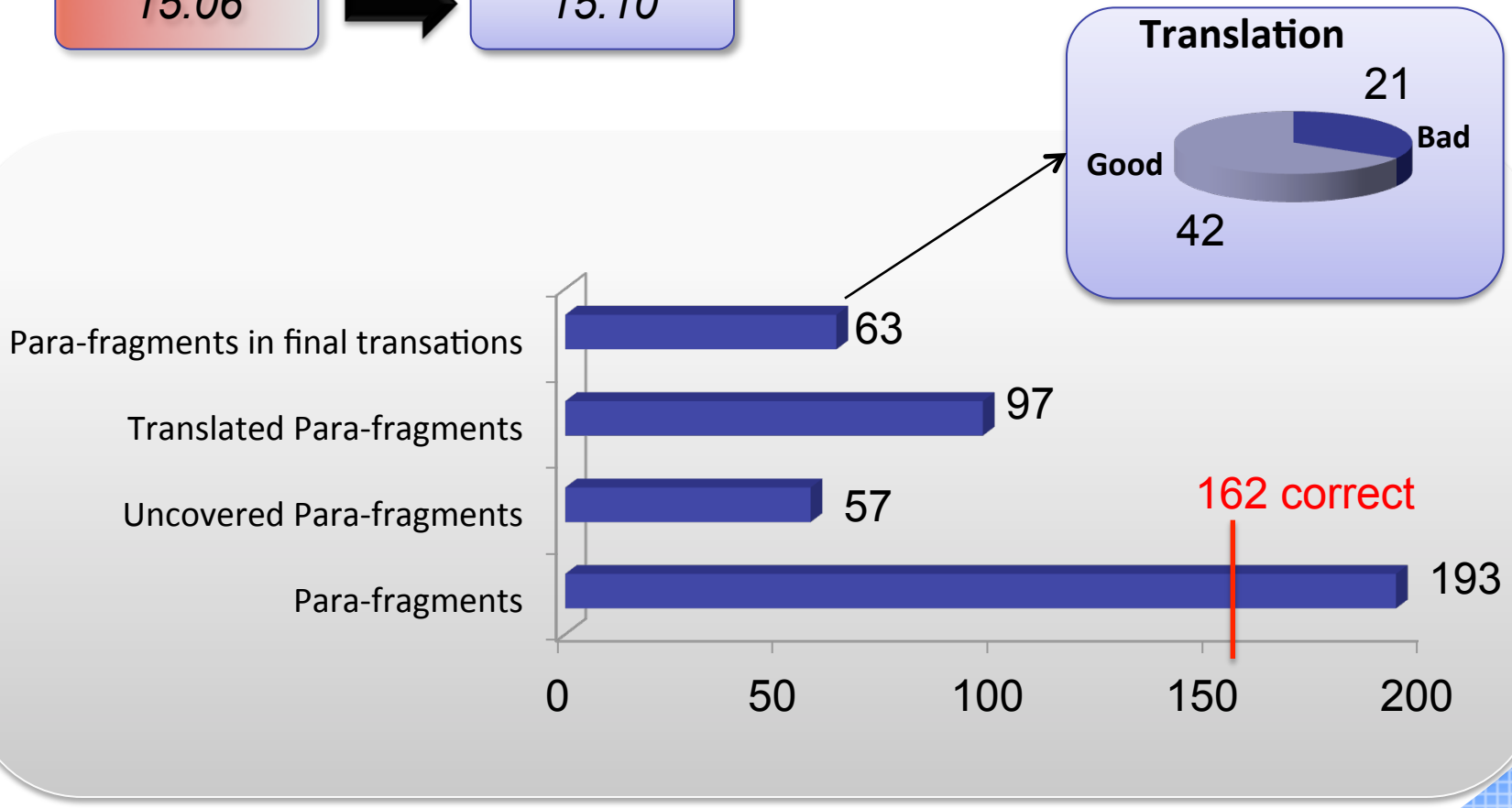
- Testing on 586 sentences (MT-EVAL 09)
- Size of parallel corpus: ~ 1.2 Million Arabic tokens

Paraphrases in Translation

Evaluation

BLEU

15.06 → 15.10



Para-fragment = a fragment based on a paraphrase-match

Conclusions

- Single-word verb paraphrases can help in matching Arabic input text with translation examples
- We believe that involving more paraphrases in translation will improve results, especially when faced with a small parallel corpus
- The paraphrasing classifier achieves a relatively high precision but maybe lower recall. Other techniques should be tested

Future Work

- Add more features to the paraphrase extraction process (e.g. “sibling” content words)
- Consider using a dependency syntax parser in the paraphrase extraction process
- Find multi-word Arabic paraphrases and use them in translation

Thank you

شكرا