# MorphTagger: HMM-Based Arabic Segmentation for Statistical Machine Translation

Saab Mansour
Human Language Technology and Pattern Recognition, RWTH Aachen University

## Motivation

- Arabic segmentation improves statistical MT results

| سأخبرها | بأن | تتصل |
|---|---|---|
| s>xbrhA | b>n | ttsl |

Segmentation

| س | أخبر | ها | ب | أن | تتصل |
|---|---|---|---|---|---|
| s | >xbr | hA | b | >n | ttsl |

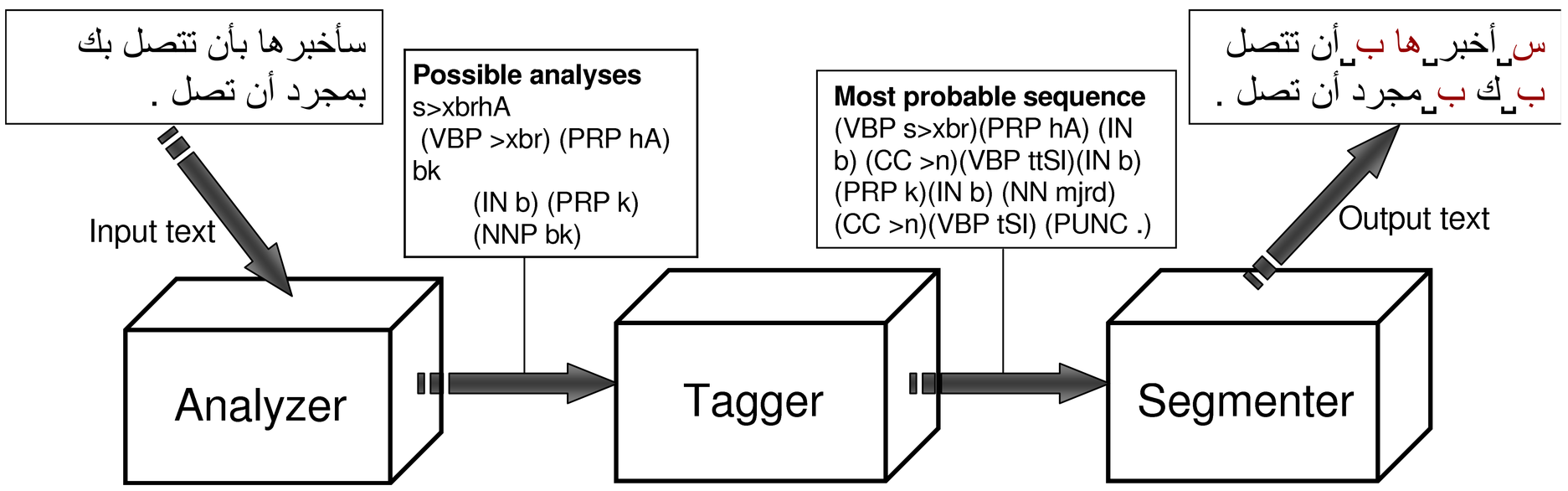
| i | 'll | tell | her | to | call |
|---|---|---|---|---|---|

- real-time SMT requires simple and fast preprocessing
- consistent comparison of segmenters

## Related work

- [Lee 04] - IBM Research
  - Ar-En. Language model-based segmentation selection
- [Habash and Sadat 06] - Columbia University (MADA)
  - Ar-En. SVM based, selects from Buckwalter Analyzer
- [El Isbihani & Khadivi+ 06] - RWTH Aachen
  - Ar-En, Ar-Fr. Rule based (FST) segmentation (no context)
- showed improvements for small tasks, but none or inferior results for big tasks (with respect to the baseline)

## MorphTagger Architecture




## Implementation

- **Morphological Analyzer**
  - Buckwalter Arabic Morphological Analyzer (BAMA) v1.0
  - rule based analyzer, with 80 000 lexicon entries
- **Tagger (POS Model)**
  - standard Markov Model Tagger (SRILM toolkit)
  - trained over the Arabic Treebank Part 1 v3.0
  - the probabilities are estimated for segments, not words!
- **Segmenter**
  - splits prepositions, possessive and objective pronouns
    $\Rightarrow$ ATB scheme
  - normalization: segments are reverted to base form
    - 'alif maksura: yX $\Rightarrow$ Y+X
    - feminine marker: tX $\Rightarrow$ p+X
    - Arabic determiner: llX $\Rightarrow$ l+Al+X

## Speed

| | speed [w/s] |
|---|---|
| FST | 4500 |
| MADA | 70 |
| MorphTagger | 1500 |

- training and tagging are fast (linear in corpus size)
- appropriate for real-time systems
- comparable tagging/segmentation accuracy to state-of-the-art [Mansour & Sima'an+ 07]

## Model

- Words: $w_1^N = w_1, ...w_n, ...w_N$
- Analyses: $a(w_n)$, $a(w_1^N)$ (given by BAMA)
- Problem: find most probable POS tags $t_1^N$:

$$t_1^N = \underset{\tilde{t}_1^N \in a(w_1^N)}{\operatorname{argmax}} Pr(\tilde{t}_1^N | w_1^N) \quad (1)$$

- Bayes decision rule and the bigram HMM model assumptions:

$$t_1^N = \underset{\tilde{t}_1^N \in a(w_1^N)}{\operatorname{argmax}} \left\{ \prod_{n=1}^{N} \left[ p(w_n|\tilde{t}_n) \cdot p(\tilde{t}_n|\tilde{t}_{n-1}) \right] \right\} \quad (2)$$

- $\{p(t_n|t_{n-1})\}$ and $\{p(w_n|t_n)\}$ are estimated on the segment level using Maximum Likelihood Estimation (MLE) followed by smoothing techniques [Mansour & Sima'an+ 07]

## Results

AR-EN BTEC 2010: 20K sentences, 150K running words (nocase+punc)

| | IWSLT04 (dev) | | IWSLT05 | | IWSLT08 | |
|---|---|---|---|---|---|---|
| System | BLEU | TER | BLEU | TER | BLEU | TER |
| TOK | 59.0 | 27.3 | 59.3 | 26.6 | 55.5 | 30.1 |
| FST | 56.4 | 29.6 | 60.2 | 26.6 | 54.5 | 30.7 |
| MorphTagger | 60.3 | 26.7 | 61.6 | 25.8 | 56.5 | 29.2 |
| MADA ATB | 59.6 | 27.5 | 60.2 | 26.6 | 55.1 | 29.5 |

AR-EN NIST 2009 subset (excludes UN and ISI): 300K sentences, 5M running words (case+punc)

| | nist04 | | nist05 | | nist06 (dev) | | nist08 | |
|---|---|---|---|---|---|---|---|---|
| System | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER |
| TOK | 47.4 | 47.3 | 49.5 | 44.9 | 39.4 | 54.1 | 36.5 | 54.3 |
| FST | 50.9 | 44.2 | 52.2 | 42.5 | 40.4 | 52.3 | 37.5 | 53.2 |
| MorphTagger | 50.6 | 44.0 | 51.9 | 42.2 | 41.5 | 51.2 | 39.7 | 51.8 |
| MADA ATB | 51.0 | 43.7 | 52.6 | 41.8 | 42.7 | 50.6 | 40.0 | 51.3 |

Lehrstuhl für Informatik 6 – Computer Science Department – RWTH Aachen University – Aachen, Germany
www: http://www-i6.informatik.rwth-aachen.de
e-mail: <mansour>@cs.rwth-aachen.de