Unsupervised False Friend Disambiguation

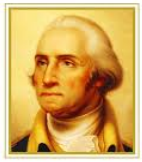
Using Contextual Word Clusters and Automatic Word Alignments

Maryam Aminian, Mahmoud Ghoneim, Mona Diab

CARE4Lang
George Washington University

SSST-9

CARE4Lang
transforming text into meaning



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

False Friend [Faux amis]

(Mitkov+ 2008)

Similar spelling

Different meaning

False Friend [Faux amis]

(Mitkov+ 2008)

Similar spelling

Different meaning

Language 1	Language 2	Similar Spelling	Different meaning	False Friend
color (En)	color (Sp)	✔	✗	NO

False Friend [Faux amis]

(Mitkov+ 2008)

Similar spelling

Different meaning

Language 1	Language 2	Similar Spelling	Different meaning	False Friend
color (En)	color (Sp)	✔	✗	NO
Library (En)	Librairie (Fr) (bookshop)	✗	✔	YES

False Friend [Faux amis]

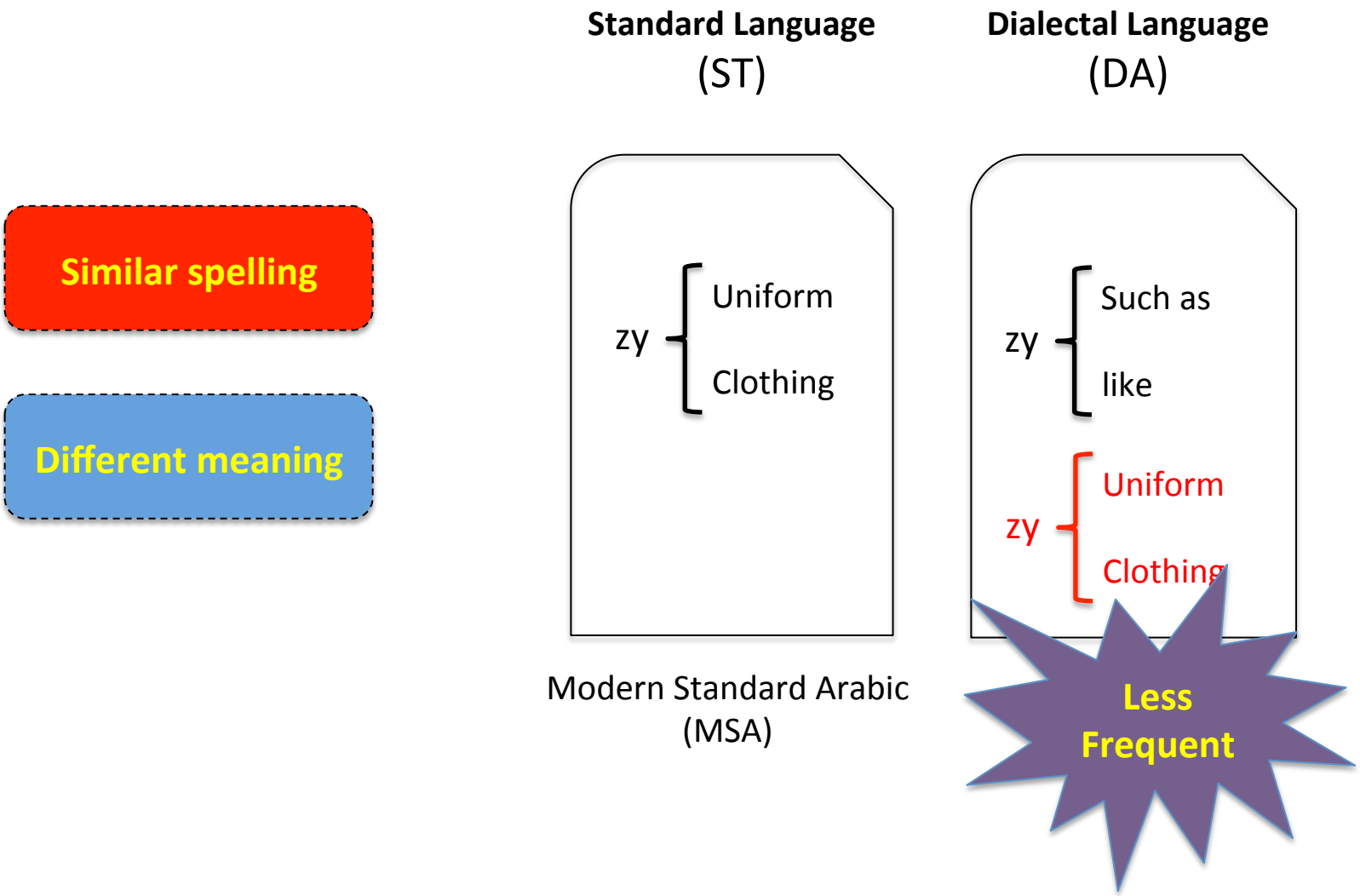
(Mitkov+ 2008)

Similar spelling

Different meaning

Language 1	Language 2	Similar Spelling	Different meaning	False Friend
color (En)	color (Sp)	✔	✗	NO
Library (En)	Librairie (Fr) (bookshop)	✔	✔	YES
Gift (En)	Gift (Gr) (poison)	✔	✔	YES

False Friend in Cross-Lang Variant Context



Background: Arabic DA vs. ST

- DAs have no standard orthographies
- DAs permeate social media
- Code switching between ST and DA within the same utterance is pervasive
- Numerous NLP tools exist for ST
- However, DA and ST variants of Arabic are significantly different on all levels of linguistic representation ***hampering direct application*** of ST NLP tools to DA processing

In MT context: Motivating Example

Egyptian: mc mlkyp xASp yEny AqSd zy AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx

Reference: not private , I mean like buses and the metro and trains ... etc .

In MT context: Motivating Example

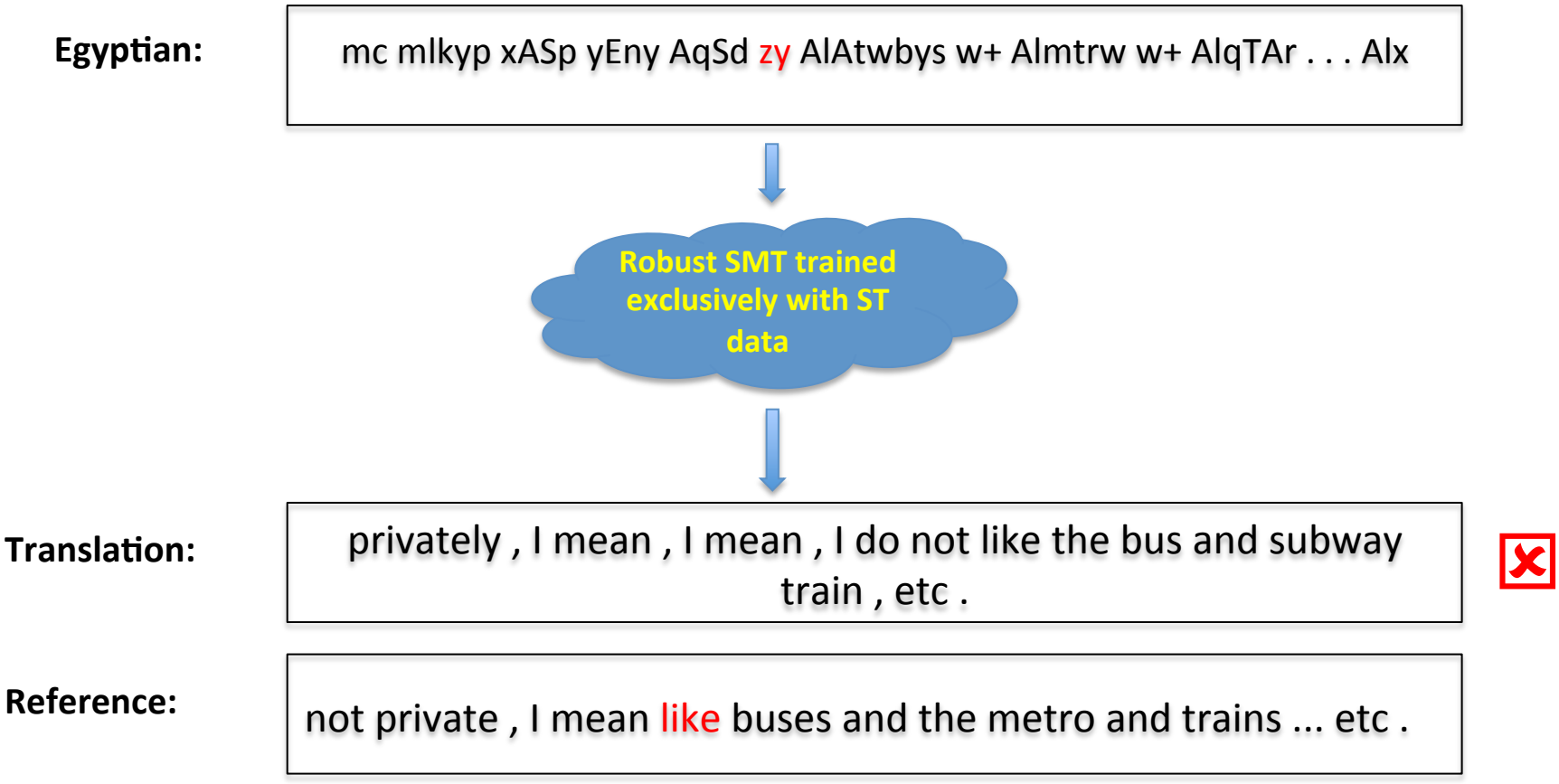
Egyptian: mc mlkyp xASp yEny AqSd zy AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx

Not enough DA parallel data to train the translation model and build stand alone machine translation systems for DA

Robust SMT systems exist for ST

Reference: not private , I mean like buses and the metro and trains ... etc .

In MT context: Motivating Example



In MT context: Motivating Example

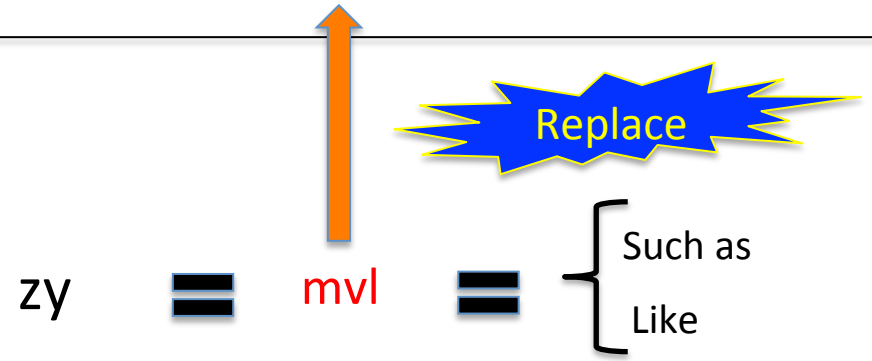
Egyptian: mc mlkyp xASp yEny AqSd zy AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx

zy = mvl = Such as / Like

Reference: not private , I mean like buses and the metro and trains ... etc .

In MT context: Motivating Example

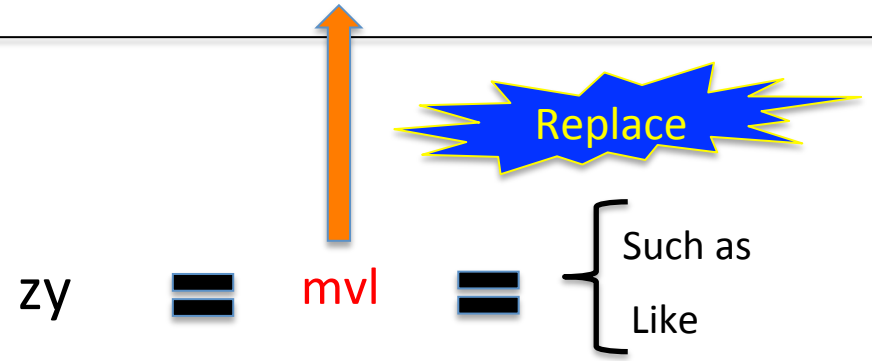
Egyptian: mc mlkyp xASp yEny AqSd zy AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx



Reference: not private , I mean like buses and the metro and trains ... etc .

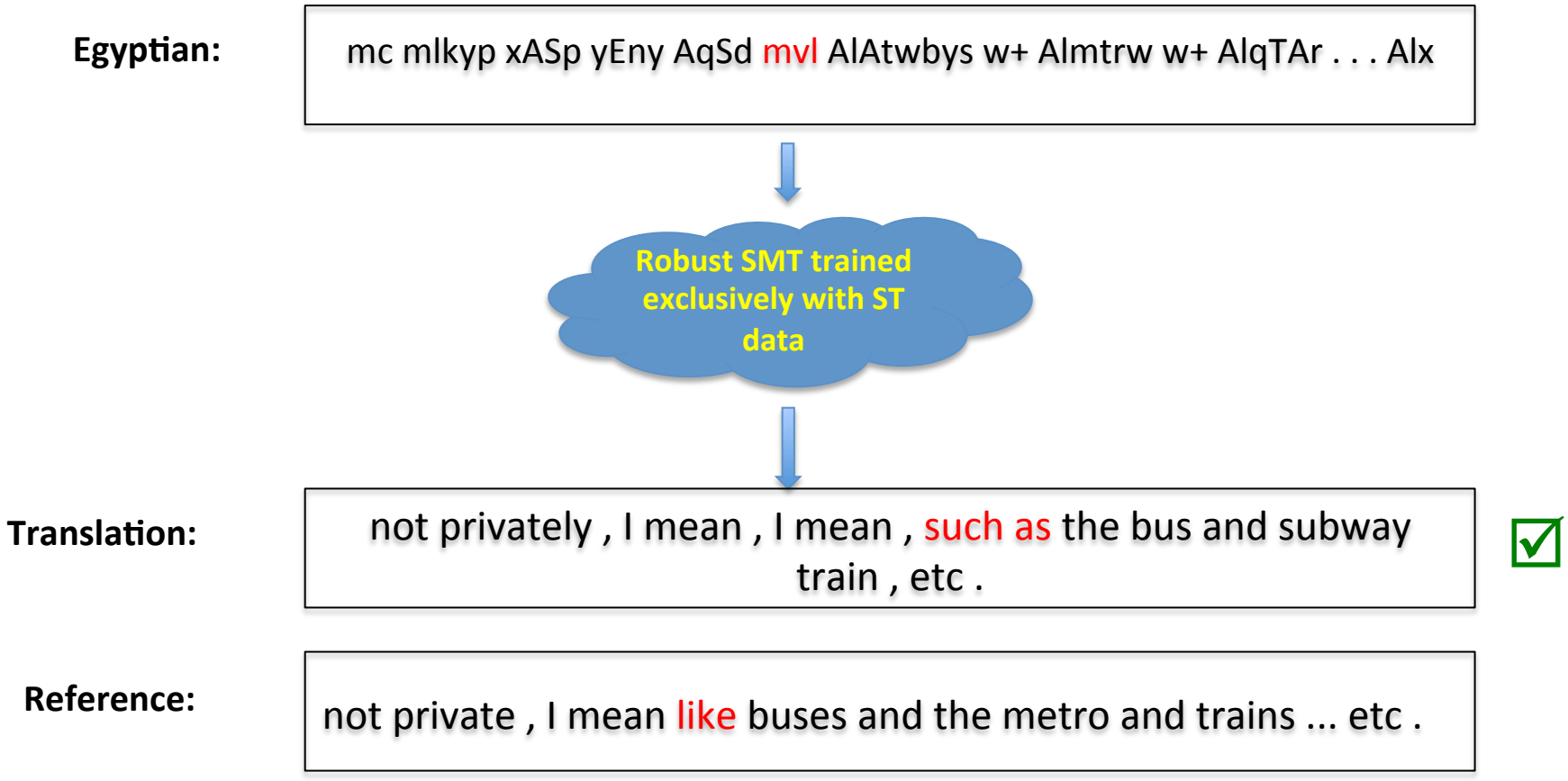
In MT context: Motivating Example

Egyptian: mc mlkyp xASp yEny AqSd mvl AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx



Reference: not private , I mean like buses and the metro and trains ... etc .

In MT context: Motivating Example

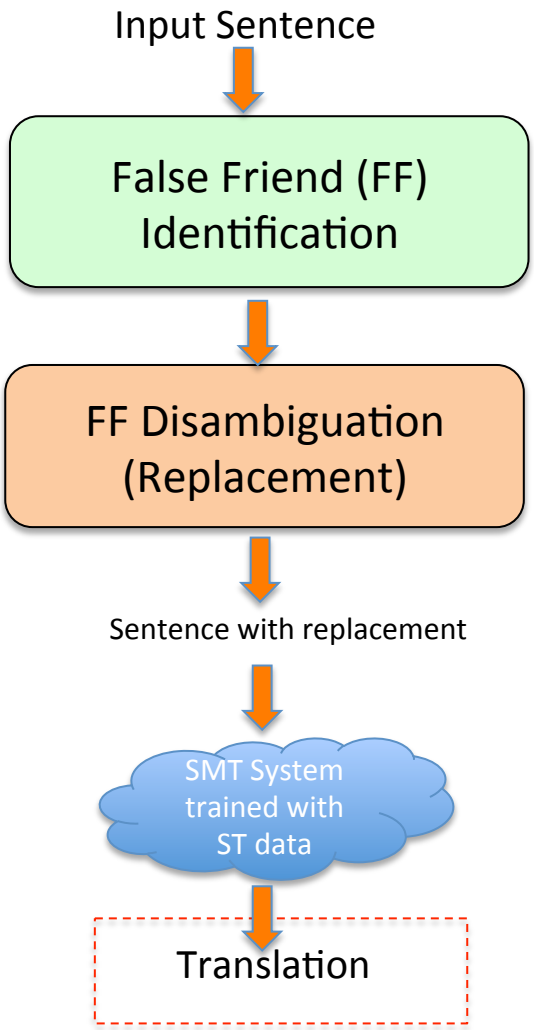


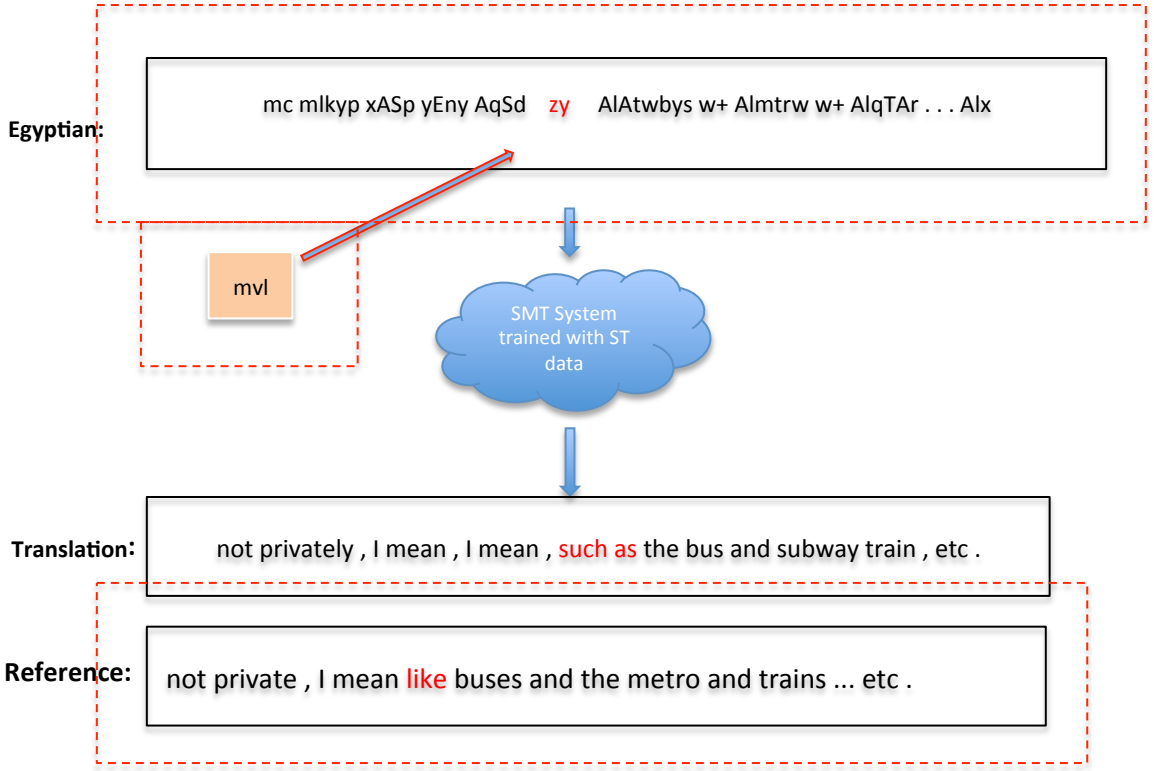
Our Goal

Enhance cross- language variant SMT performance,
crucially, in absence of
in-domain training data

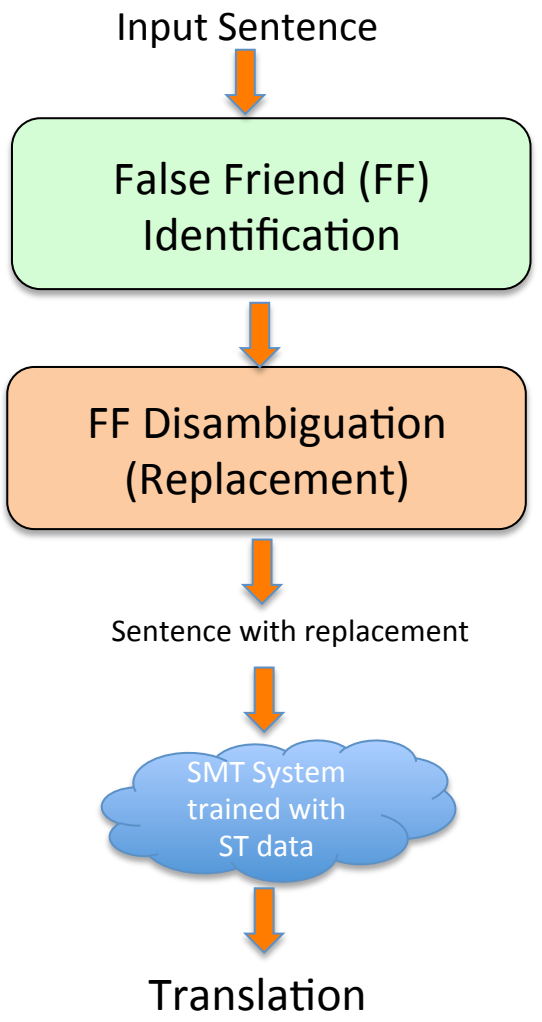
i.e. using an *exclusively* ST system to translate DA data

Our Approach





Our Approach



✓ Er… …ord ali… …ARL) corpus

PARL Classifier

✓ Re… …c equivalent fr…

WC Classifier

✓ Us… …clusters (WC) to model the context

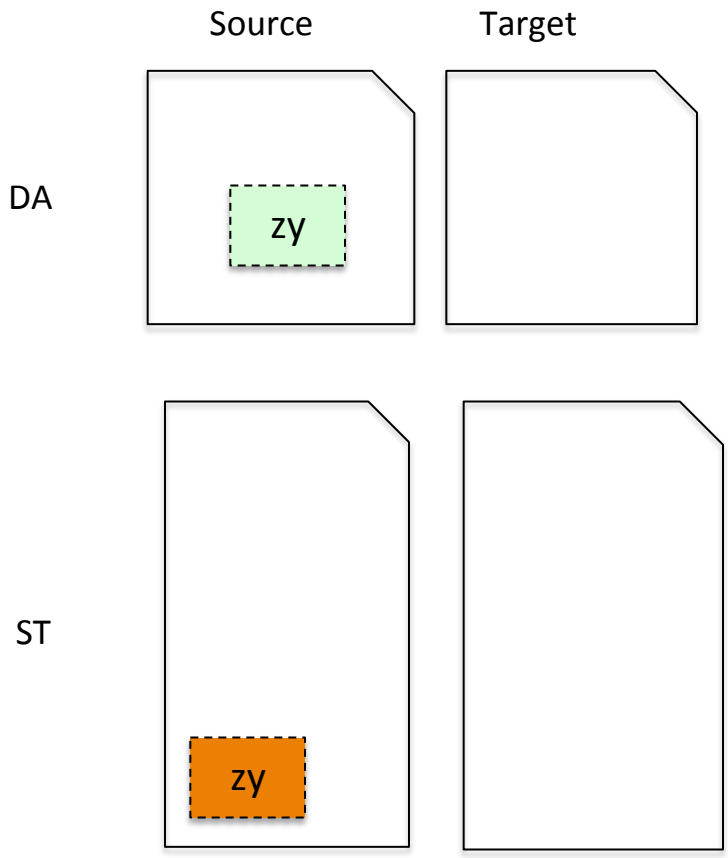
Extrinsic Evaluation

PARL Classifier

- ✓ There is no labeled data with FF tags
- ✓ Training data for PARL is created automatically

PARL Classifier

Generating Training Data

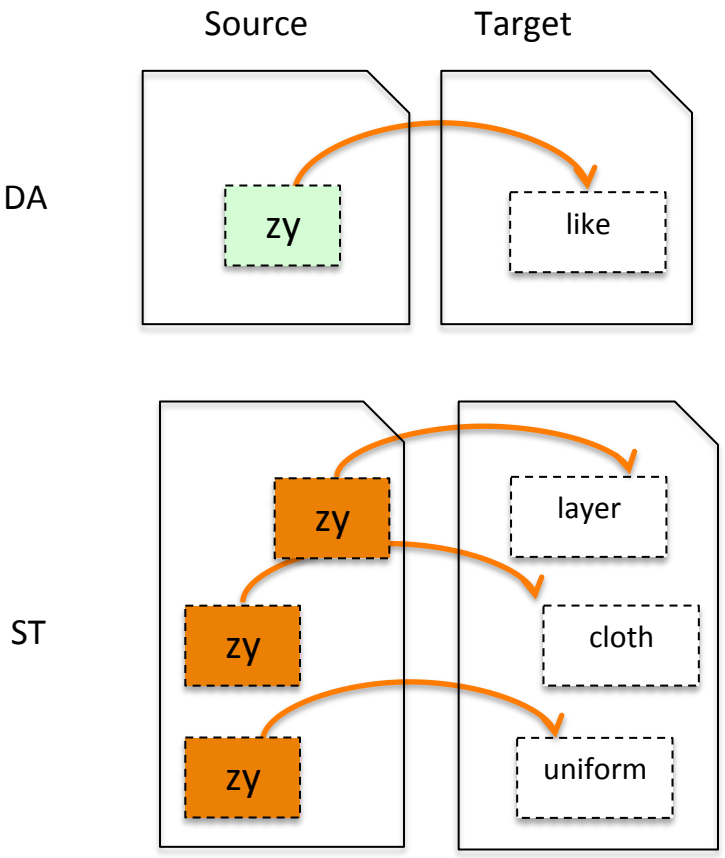


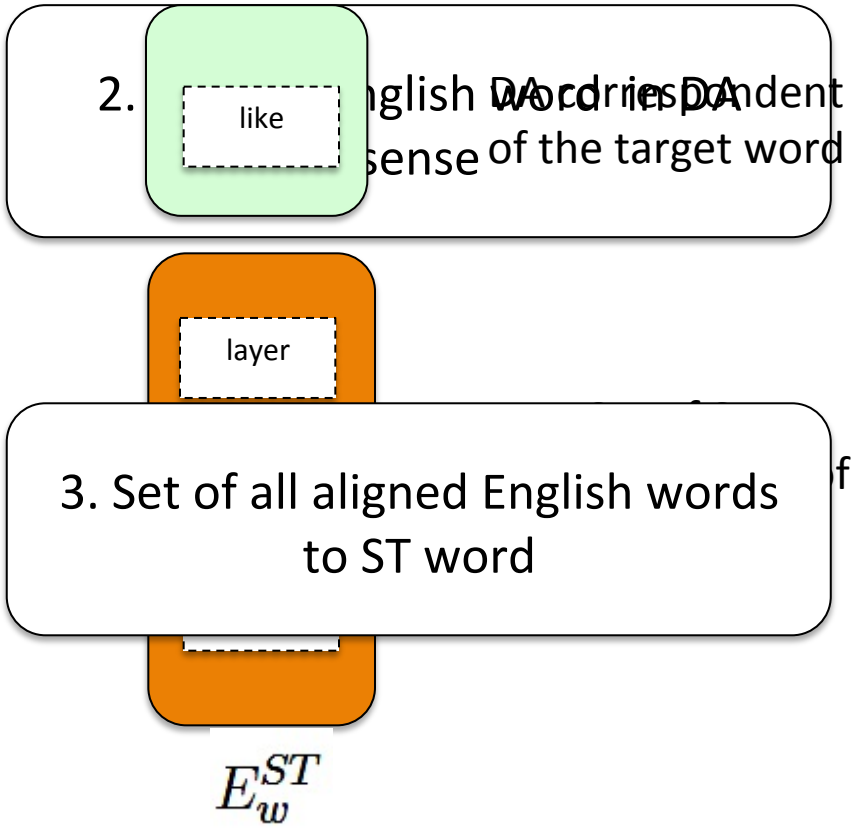
1. Identify words used in both DA & ST (Cross-variant homographs)

$$\mathcal{L}(k,i) = \{\text{DA}, \text{ST}\}$$

PARL Classifier

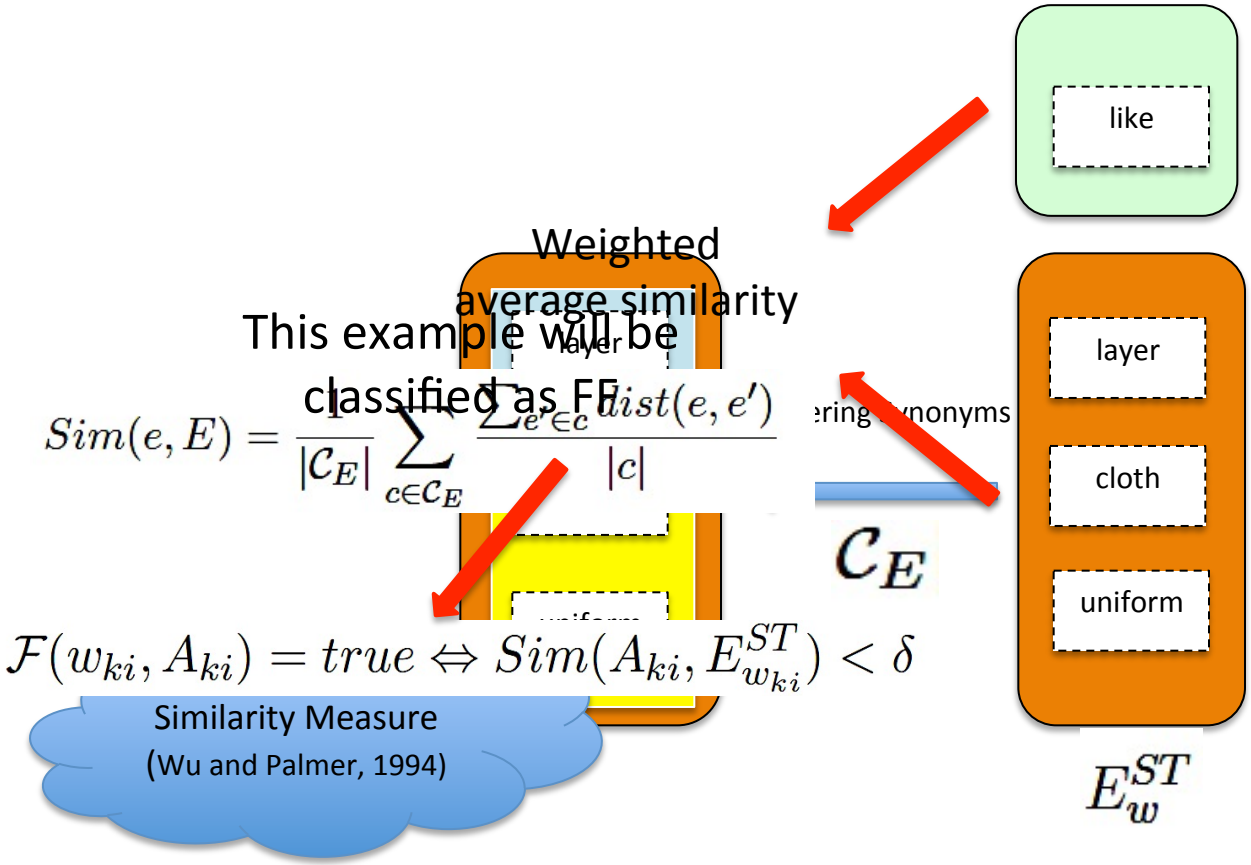
Generating Training Data





PARL Classifier

Generating Training Data



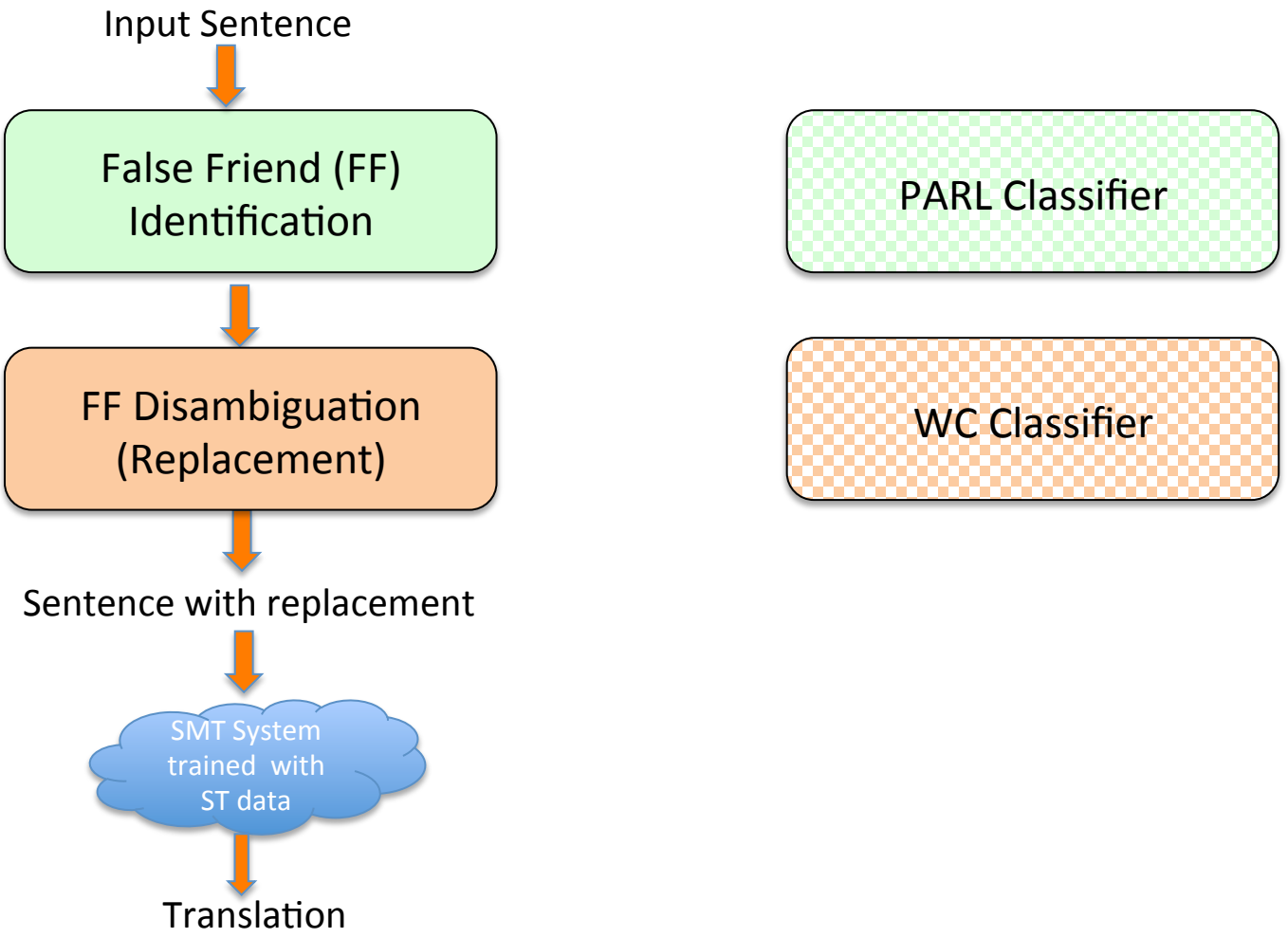
correspondent of DA word

Set of correspondents of ST word

PARL Setup

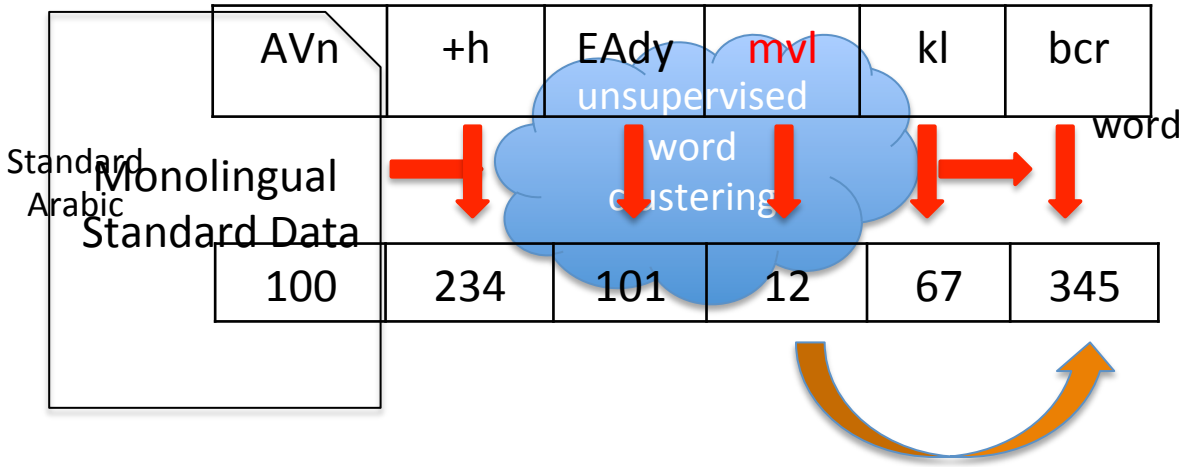
- ✓ **Averaged** Perceptron for classification
- ✓ Words represented with the following features:
  - ❖ Lemma of current word
  - ❖ POS of current word
  - ❖ POS of previous word
  - ❖ POS of next word

Our Approach



WC Classifier

Training



$$P_\tau(c|w) \text{ for } \tau \in \{-2, -1, +1, +2\}$$

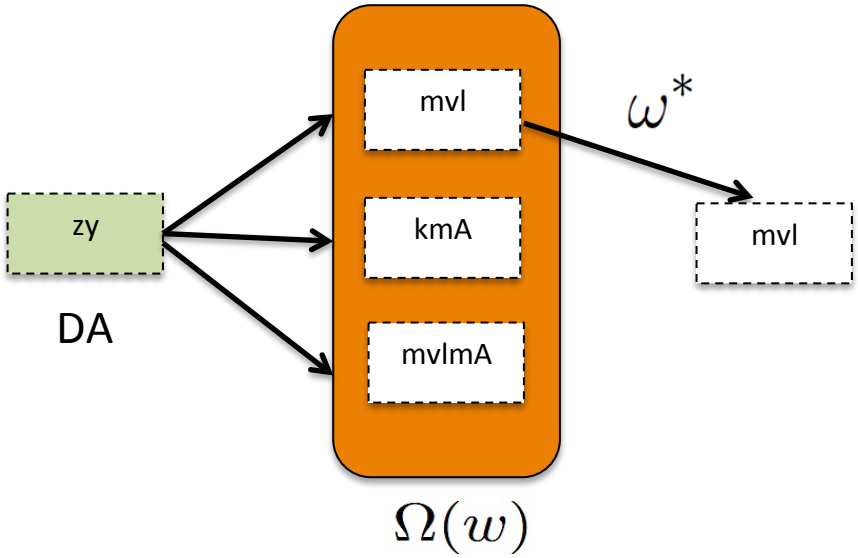
$$c \in \{1, 2, ..., K\}$$

Estimated using maximum likelihood estimation with additive smoothing

WC Classifier

Disambiguation

✓ Given a set of predefined ST equivalents for each DA word w:



mc mlkyp xASp yEny AqSd **zy**
AlAtwbys w+ Almtrw w+
WC AlqTAr . . . Alx

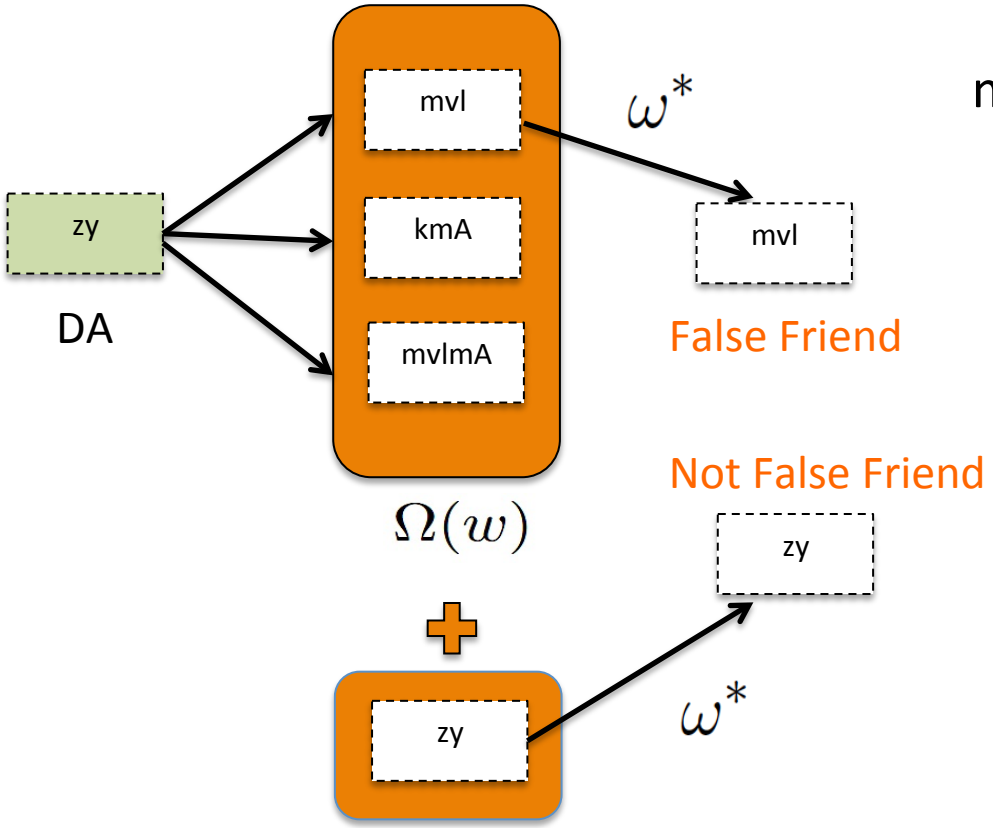
$$\omega^* = \quad \text{...} \quad c_T|\omega)$$

Hypothesis: This ST equivalent is more likely to appear in the context compared to other possible equivalents

WC Classifier

Disambiguation

✓ Given a set of predefined ST equivalents for each DA word w:



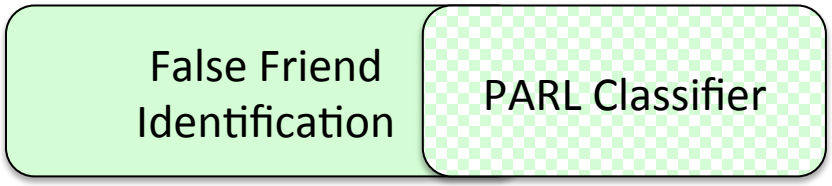
mc mlkyp xASp yEny AqSd **zy**
AlAtwbys w+ Almtrw w+
AlqTAr . . . Alx

WC_{cor}

EAdAt ylbs **zy** cEby

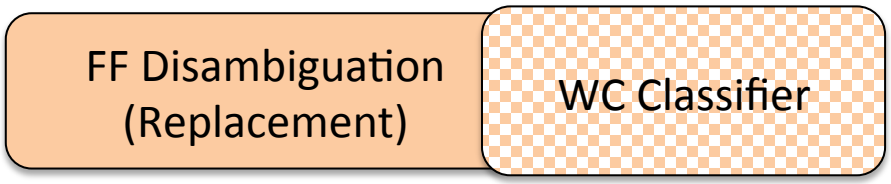
Experimental Setup

- ✓ Trained using ST parallel data from multiple LDC corpora
- ✓ GIZA++ (Och and Ney, 2003) for word alignment
- ✓ AIDA: Token dialect identification tool (Elfardy et al., 2013)
  - ❖ AIDA provides a list of ST equivalents for each DA word



Experimental Setup

- ✓ Trained using Arabic Gigaword 4
- ✓ word2vec (Mikolov et al., 2013) K-means word clustering tool to obtain word clusters



Evaluation

- ✓ Extrinsic evaluation of FF using SMT system
- ✓ Evaluation metrics: BLEU, METEOR, TER, WER, PER
- ✓ Evaluation set: BOLT-ARZ DA data set
- ✓ SMT setup:
  - ❖ Moses decoder to build a standard phrase-based SMT system
  - ❖ Factored translation model with lemma and POS factors
  - ❖ Feature weights are tuned to maximize BLEU score on the tuning set using MERT
  - ❖ Final results are reported by averaging over 3 tuning sessions with random initialization
  - ❖ SRILM to build 5-gram language models with modified Kneser-Ney smoothing

Experimental Conditions

Baselines

	FF Identification	FF Replacement
No Replacement Baseline	✗	✗
Random Baseline	DA-ST Homographs	Random subset of FF set Random equivalent selection
Blind Baseline	DA-ST Homographs	Whole set of FF Random equivalent selection

Experimental Conditions

		FF Identification	FF Replacement
Baselines	No Replacement Baseline	✗	✗
	Random Baseline	DA-ST Homographs	Random subset of FF set Random equivalent selection
	Blind Baseline	DA-ST Homographs	Whole set of FF Random equivalent selection
Replacement	PARL	PARL	Whole set of FF Random equivalent selection
	WC	WC	WC
	WC_{cor}	WC_{cor}	WC_{cor}
	PARL+WC	PARL	WC
	PARL+WC_{cor}	PARL	WC_{cor}

Experimental Conditions

Baselines

	FF Identification	FF Replacement
No Replacement Baseline	✗	✗
Random Baseline	DA-ST Homographs	Random subset of FF set Random equivalent selection
Blind Baseline	DA-ST Homographs	Whole set of FF Random equivalent selection

- ✓ **Random** and **Blind** Baselines contrast impact of PARL on SMT performance
- ✓ **No Replacement** Baseline contrast impact of whole pipeline on SMT performance

Results

	BLEU	METEOR	TER	WER	PER
Rand. Baseline	20.6	27.5	65.9	69.2	45.3
Blind Baseline	20.1	27.2	68.3	71.6	46.6
PARL	20.7	27.1	67.5	69.6	45.5

- ✓ Using PARL for FF identification improves SMT performance compared to Random and Blind baselines

Results

	BLEU	METEOR	TER	WER	PER
NoReplac. Baseline	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC	20.8	27.5	66.2	69.1	45.2
WC_{cor}	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC_{cor}	21.3	27.9	65.5	**68.0**	**44.5**

- ✓ Using contextual word clusters for FF identification and disambiguation has a higher impact on final SMT performance compared to PARL component

Results

	BLEU	METEOR	TER	WER	PER
NoReplac. Baseline	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC	20.8	27.5	66.2	69.1	45.2
WC_{cor}	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC_{cor}	21.3	27.9	65.5	**68.0**	**44.5**

- ✓ WC_{cor} disambiguation module results higher improvement in SMT performance compared to WC

Results

	BLEU	METEOR	TER	WER	PER
NoReplac. Baseline	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC	20.8	27.5	66.2	69.1	45.2
WC_{cor}	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC_{cor}	21.3	27.9	65.5	**68.0**	**44.5**

- ✓ PARL+WC_{cor} rectify some of the mistakes from PARL by using WC_{cor} as an additional FF identifier

Results

	BLEU	METEOR	TER	WER	PER
NoReplac. Baseline	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC	20.8	27.5	66.2	69.1	45.2
WC_{cor}	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC_{cor}	21.3	27.9	65.5	**68.0**	**44.5**

- ✓ Though no improvement over vanilla baseline in bleu scores, our approach has the power to enhance SMT lexical choice and select more accurate target translations for the false friends.

Results

	BLEU	METEOR	TER	WER	PER
NoReplac. Baseline	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC	20.8	27.5	66.2	69.1	45.2
WC_{cor}	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC_{cor}	21.3	27.9	65.5	**68.0**	**44.5**

- ✓ Issue is that the SMT translation table does not contain adequate bilingual phrase pairs for some of the replaced MSA equivalents (suggested by AIDA tool)

Error Analysis

Ref.	let us forget about our differences and unite .
Input DA	**nsyb** +nA mn AlAxtlAf w+ ntwHd
Sentence with Replacement	**trk** +nA mn AlAxtlAf w+ ntwHd
Baseline Trans.	we disagree and suffering from
Replacement Trans.	let us *leave* the difference and unify

- ✓ Word 'nsyb' which means *forget* in this context is replaced with MSA equivalent 'trk' that means *leave* or *forget*
- ✓ Decoder has translated phrase 'trk +nA mn AlAxt- lAf' into a longer phrase *let us leave the difference* instead of generating an incoherent translation such as baseline

Error Analysis

Ref.	and those who said that the girls ... indeed , i heard very bad words , why ?
Input DA	w+ Ally yqwl AlbnAt . . . b+ jd smEt AllfAZ **wHcp** qwy lyh kdh
Sentence with Replacement	w+ Ally yqwl AlbnAt . . . b+ jd smEt AllfAZ **syC** qwy lyh kdh
Baseline Trans.	and to say ... very very difficult . that is why i heard
Replacement Trans.	and to say ... seriously , i heard a *strong bad* words , why ?

- ✓ Word 'wHcp' in the third example is not a pure EGY word
- ✓ However, it conveys a meaning different from its observed senses in the phrase table (meaning "to miss someone or difficult")
- ✓ Our approach has improved SMT lexical choice significantly in this example

Error Analysis

Ref.	also eradication of poverty and need is very important , toqua
Input DA	w+ kmAn AlqDAC Ely Alfqr w+ **HAjp** mhm jdA yA+ tqy
Sentence with Replacement	w+ kmAn AlqDAC Ely Alfqr w+ **Amr** kbyr jdA yA+ tqy
Baseline Trans.	and also the eradication of poverty and need is very important , and to say ... very very difficult . that is why i heard
Replacement Trans.	and also the eradication of poverty and a very large ,

- ✓ FF identifier has incorrectly identified word 'HAjp' (*need* in this context) as FF
- ✓ Decoder is not able to find a proper translation for the replaced word in the context

Error Analysis

Ref.	i will tell you a story , and you judge whose fault it is .
Input DA	**Tb** AnA H+ AHky l+ HDrp +k mwqf w+ tqwly myn Ally glTAn
Sentence with Replacement	**tmAm** AnA H+ AHky l+ HDrp +k mwqf w+ tqwly myn Ally glTAn
Baseline Trans.	ok , I am going to talk to you and say who was wrong .
Replacement Trans.	I will talk to you stand and say who was wrong .

- ✓ Word '*Tb*' in the baseline sentence means *all right*, *very well* or *ok* in EGY while it means *medicine* when used in MSA
- ✓ FF identifier has correctly identified this word as a FF.
- ✓ WC disambiguator module also has adequately replaced word 'Tb' with the MSA word 'tmAm'

Contributions

- ✓ We presented a new approach for improving cross-dialect SMT performance without ***any*** in-domain training data
- ✓ We showed that our approach improves DA-EN SMT lexical choice
- ✓ We devised an unsupervised effective approach for false friend identification and disambiguation

Future Work

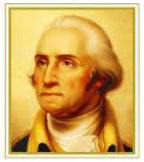
- ✓ Exploring an automatic way to generate the list of possible equivalents for FF
- ✓ Benefiting from continuous word vectors and their similarity to extract possible word senses for a particular FF

Thank you!

Special thanks to
QTLeap best paper award
committee

http://www.seas.gwu.edu/~aminian/

CARE4Lang
transforming text into meaning



THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON DC