CS-AD 220 – Spring 2016

Natural Language Processing

Session 24: 26-Apr-16

Prof. Nizar Habash

NYUAD Course CS-AD 220 – Spring 2016

Natural Language Processing

Assignment #4

Phrase-based Statistical Machine Translation

Assigned Apr 19, 2016

Due May 10, 2016 (11:59pm)

Introduction[1]

In this laboratory exercise, you will build a complete phrase-based statistical machine translation system from small amounts of training data, evaluate their performance, and identify ways that translation quality can be improved. Resulting systems will be evaluated on test data (released a few days before the deadline). You will build the MT system using Moses, an open-source phrase-based statistical machine translation decoder.

Assignment #4 posted on NYU Classes

START EARLY!

DEADLINE IS May 10 (11:59pm)

Challenges for Statistical MT

- Data Sparsity
  - Training models need a lot of data
  - Genre and domain sensitive
  - Worse for language with rich morphology
  - Some language pairs have little to no parallel data

Challenges for Statistical MT

- Data Sparsity
  - Training models need a lot of data
  - Genre and domain sensitive
  - **Worse for language with rich morphology**
  - Some language pairs have little to no parallel data
- Solution: Tokenization
  - Break up the words to symmetrize source and target languages
    - Arabic wsyktbhA And he will write it
    - ➔ w+ s+ yktb +hA

Arabic, English and MT

	Arabic	English
Orthographic ambiguity	More	Less
Orthographic inconsistency	More	Less
Morphological inflections	More	Less
Morpho-syntactic complexity	More	Less
Word order freedom	More	Less
Dialectal variations	More	Less

Road Map

- Machine Translation for Arabic
  - **Tokenization for Arabic to English MT**
  - OOV Reduction
  - Dialect to English MT through MSA Pivoting

Preprocessing Schemes

Input:	wsyktbhA?	‘and he will write it?’

Preprocessing Schemes

- ST Simple Tokenization

Input:		wsyktbhA?	'and he will write it?'
	ST	wsyktbhA ?	

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+

Input:		wsyktbhA?	‘and he will write it?’
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+
- D2 Decliticize CONJ+, PART+

Input:		wsyktbhA?	'and he will write it?'
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+
- D2 Decliticize CONJ+, PART+
- D3 Decliticize all clitics

Input:		wsyktbhA?	‘and he will write it?’
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	
	D3	w+ s+ yktb +hA ?	

Preprocessing Schemes

• ST	Simple Tokenization
• D1	Decliticize CONJ+
• D2	Decliticize CONJ+, PART+
• D3	Decliticize all clitics
• BW	Morphological stem and affixes

Input:		wsyktbhA?	'and he will write it?'
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	
	D3	w+ s+ yktb +hA ?	
	BW	w+ s+ y+ ktb +hA ?	

(Habash&Sadat, 2006)

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+
- D2 Decliticize CONJ+, PART+
- D3 Decliticize all clitics
- BW Morphological stem and affixes
- EN D3, Lemmatize, English-like POS tags, Subj

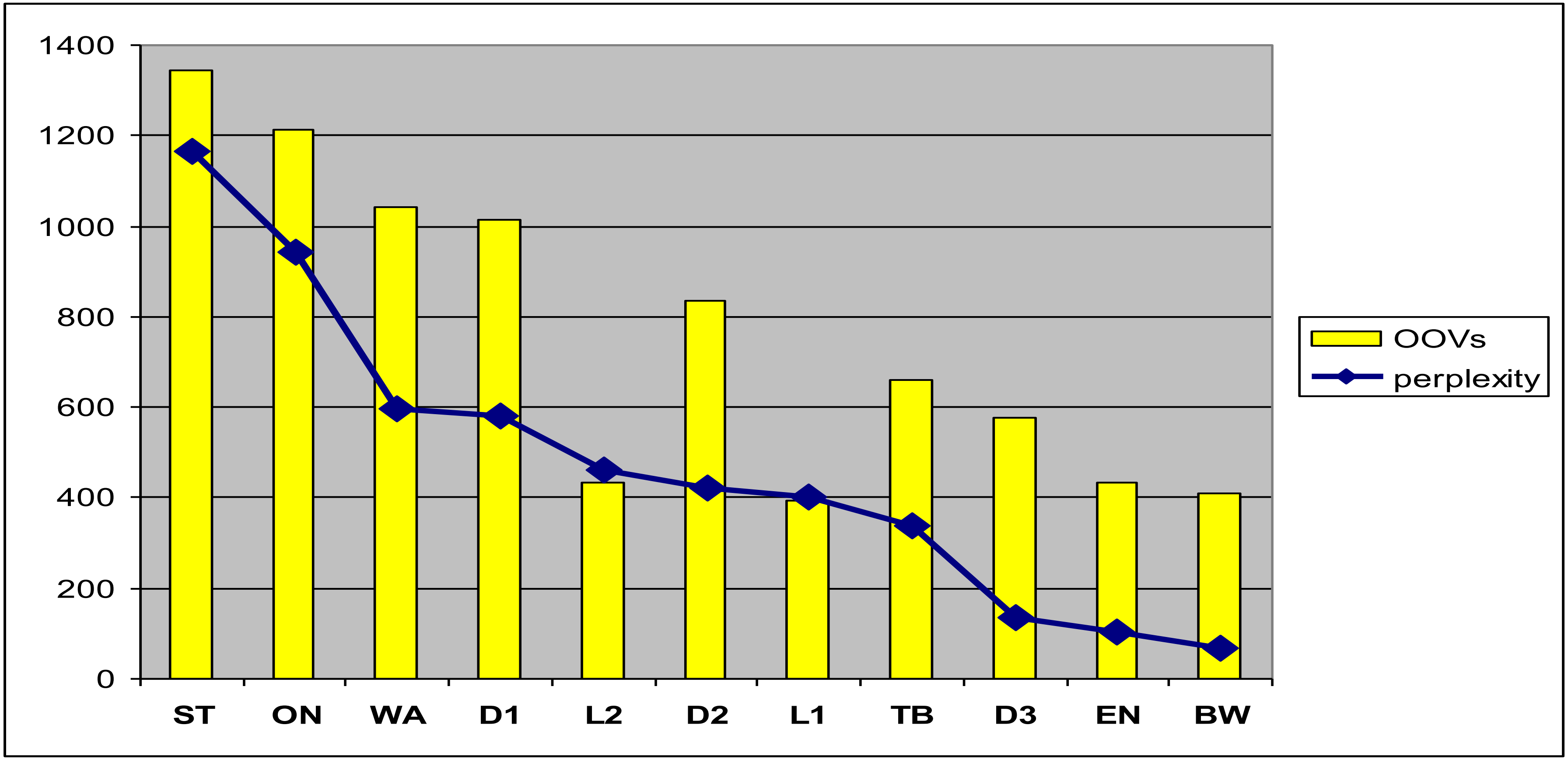
Input:		wsyktbhA?	‘and he will write it?’
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	
	D3	w+ s+ yktb +hA ?	
	BW	w+ s+ y+ ktb +hA ?	
	EN	w+ s+ ktb/VBZ S:3MS +hA ?	

Preprocessing Schemes

• ST	Simple Tokenization
• D1	Decliticize CONJ+
• D2	Decliticize CONJ+, PART+
• D3	Decliticize all clitics
• BW	Morphological stem and affixes
• EN	D3, Lemmatize, English-like POS tags, Subj
• ON	Orthographic Normalization
• WA	wa+ decliticization
• TB	Arabic Treebank
• L1	Lemmatize, Arabic POS tags
• L2	Lemmatize, English-like POS tags

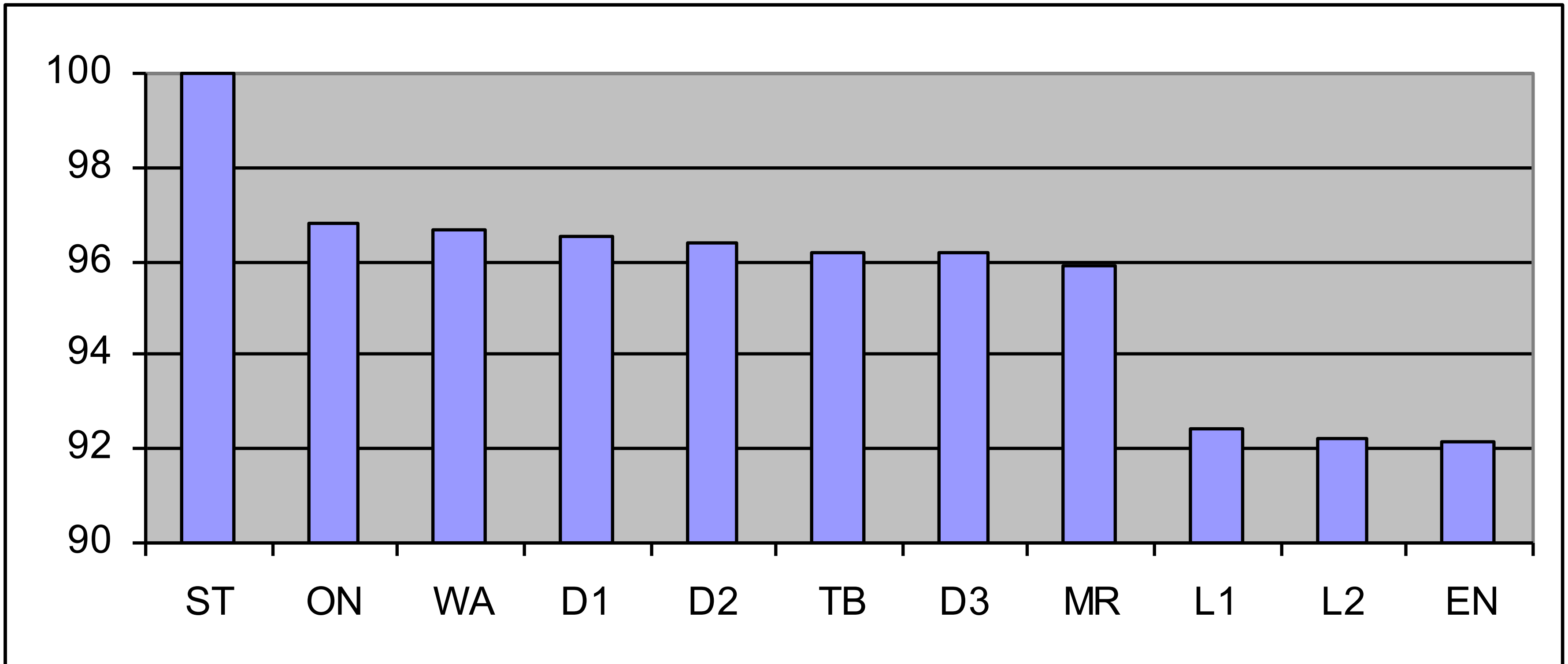
Preprocessing Schemes

- OOVs and Perplexity
  - MT04,1353 sentences, 36000 words



Preprocessing Schemes

- Scheme Accuracy
  - Measured against Penn Arabic Treebank



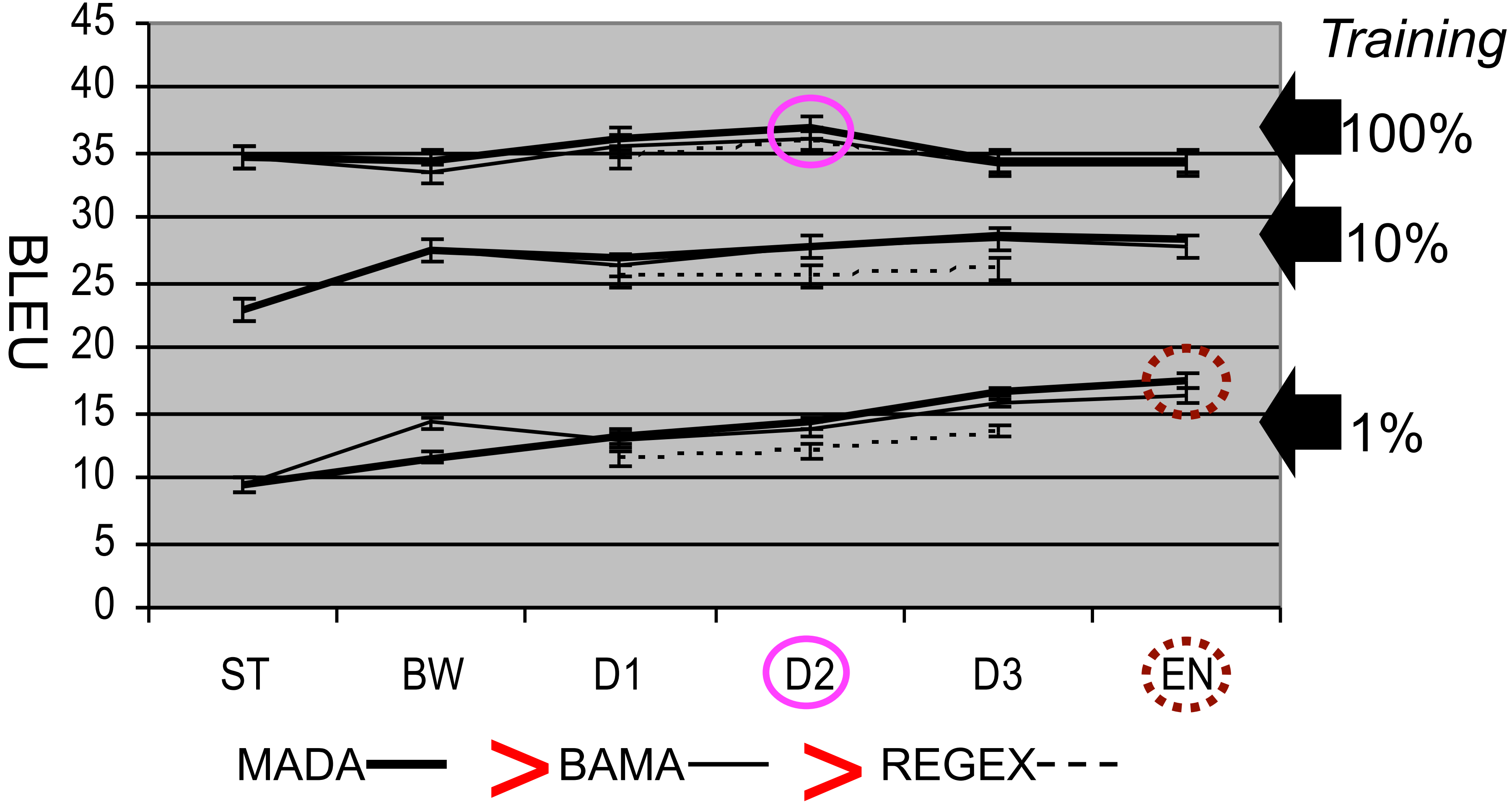
Preprocessing Techniques

- **REGEX**: Regular Expressions
- **BAMA**: Buckwalter Arabic Morphological Analyzer (Buckwalter 2002; 2004)
  - Pick first analysis
  - Use TOKAN (Habash, 2007)
- **MADA**: *Morphological Analysis and Disambiguation for Arabic* (Habash&Rambow, 2005)
  - Multiple SVM classifiers + combiner
  - Selects BAMA analysis
  - Use TOKAN

Experiments

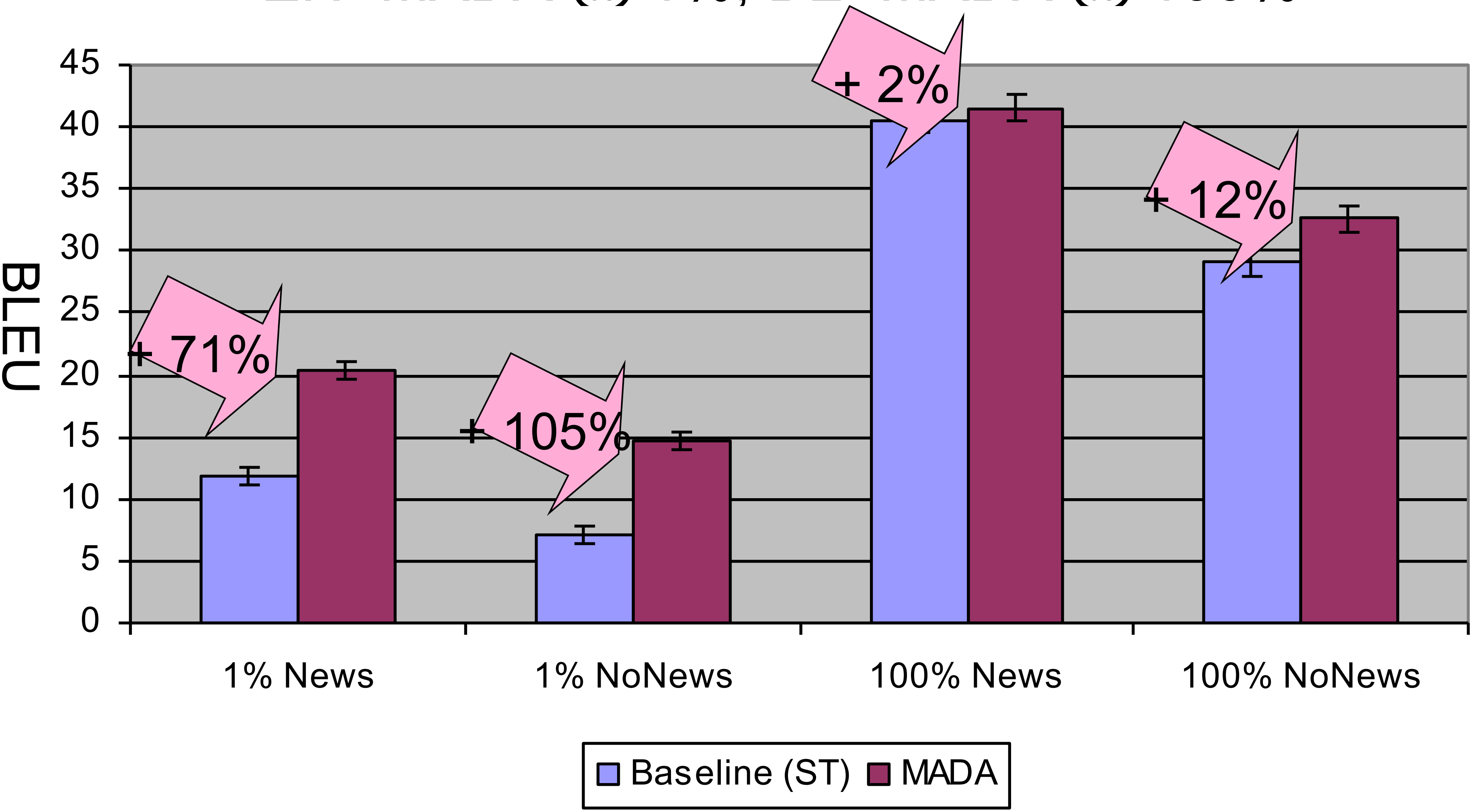
- Portage Phrase-based MT (Sadat et al., 2005)
- Training Data: parallel 5 Million words only
  - All in News genre; Learning curve: 1%, 10% and 100%
- Language Modeling: 250 Million words
- Development Tuning Data: MT03 Eval Set
- Test Data MT04
  - Mixed genre: news, speeches, editorials
- Each experiment
  - Select a preprocessing scheme
  - Select a preprocessing technique
  - Some combinations do not exist, e.g., REGEX and EN

MT04 Results



MT04 Genre Variation

Best Schemes + Technique
EN+MADA @ 1%, D2+MADA @ 100%



Lessons Learned

- For large amounts of training data, splitting off conjunctions and particles performs best
- For small amount of training data, following an English-like tokenization performs best
- Suitable choice of preprocessing scheme and technique yields an important increase in BLEU score if
  - there is little training data
  - there is a change in genre between training and test
- Using MADA+TOKAN provides a framework no more.
- Differences in MT approach, data genre and size require the developers to study the behavior under different settings.
  - For Phrase-based MT, D2/ATB does best; Other approaches do better with D3

Road Map

- Machine Translation for Arabic
  - Tokenization for Arabic to English MT
  - **OOV Reduction**
  - Dialect to English MT through MSA Pivoting

(Habash, 2008)
http://www.aclweb.org/anthology/P08-2015

REMOOV

- Out-Of-Vocabulary (OOV)
  - Test words that are not modeled in training
  - May be in training data but not in phrase table
  - May be in phrase table but not matchable
- A persistent problem
  - Arabic in ATB tokenization with orthographic normalization:

  Increasing the training data by 12 times

  → 66% reduction in Token/Type OOV

  → 55% reduction in Sentence OOV (sentences with at least 1 OOV word)

	Medium			**Large**		
Word count	**4.1M**			**47M**		
	MT03	**MT 04**	**MT 05**	**MT03**	**MT 04**	**MT 05**
Token OOV	**2.5%**	**3.2%**	**3.0%**	**0.8%**	**1.1%**	**1.1%**
Type OOV	**8.4%**	**13.32%**	**11.4%**	**2.7%**	**4.6%**	**4.0%**
Sentence OOV	**40.1%**	**54.47%**	**48.3%**	**16.9%**	**25.6%**	**22.8%**

Profile of OOVs in Arabic

- Proper nouns (40%)
  - Different origins: Arabic, Hebrew, English, French, Italian, and Chinese
- Other parts-of-speech (60%)
  - Nouns (26.4%), Verbs (19.3%) and Adjectives (14.3%)
  - Less common morphological forms such as the dual form of a noun or a verb
- Orthogonally, spelling errors appear in (6%) of cases and tokenization errors appear in (7%) of cases

Proper Noun	**40%**	**روثبين، جفعاتايم، هوكايدو**
Noun/Adjective	**41%**	**قريتين، مدرستا**
Verb	**19%**	**سيلتقيان، تر، مررنا**
Spelling Error	**13%**	**اشحاض، باكتسان، لروثبين**

OOV Reduction Techniques

- Two strategies for online handling of OOVs by phrase table extension
  - Recycle Phrases
    - Expand the phrase table online with recycled phrases
      - Relate OOV word to INV (in-vocabulary) word
      - Copy INV phrases and replace INV word with OOV word
      - Example: add misspelled variant of a word in phrase table
        knAb كناب »
      - Using unigram and bigram phrases was optimal for BLEU
  - Novel Phrases
    - Expand the phrase table online with new phrases
      - Example: باستور *bAstwr* is OOV
      - Use transliteration software to produce possible translations
        - » Pasteur, Pastor, Pastory, Bostrom, etc.

REMOOV Techniques

- MorphEx (morphological expansion)
- DictEx (dictionary expansion)
- SpellEx (spelling expansion)
- TransEx (name transliteration)

	Morphology	**No Morphology**
Recycled Phrases	*MorphEx*	*SpellEx*
Novel Phrases	*Dictex*	*TransEx*

Morphology Expansion

- Model target-irrelevant source morphological variations
  - Cluster Arabic translations of English words
    - book ← كتاب, الكتاب, كتابا
    - write ← يكتب تكتب نكتب يكتبون يكتبن سيكتبن ...
  - Learn mappings of morphological features for words sharing lexemes in the same cluster
    - [POS:V +S:3MS] == [POS:V +S:3FS]
    - [POS:N Al+ +PL] == [POS:N +PL]
    - [POS:N +DU] == [POS:N +PL]
- Map OOV word to INV word using a morphology rule:

الجماعتين → [POS:N Al+ +DU] == [POS:N +PL] → جماعات

Spelling Expansion

- Relate an OOV word to an INV word through:
  - Letter deletion mArynA → mArnA
  - Letter insertion mArynA → mAAArynA
  - Letter inversion mArynA → mAyrnA
  - Letter substitution mArynA → mAzynA
  - Substitution in Arabic was limited to 90 cases (as opposed to 1260)
    - Shape alternations ر >< ز r >< z
    - Phonological alternations س >< ص s >< S
    - Dialectal variations أ >< ق A >< q
- *No modification of the probabilities in the recycled phrases*

Transliteration Expansion

- Use a similarity metric (Freeman et al 2006) to match Arabic spelling to English spelling of proper names
  - Expand forms by mapping to Double Metaphones (Philips, 2000)
- Assign very low probabilities that are adjusted to reflect similarity metric score

المتنبي	→	MTNP	→	Al-Mutannabi Al-Mutanabi
باستور	→	PSTR	→	Pasteur Pastor Pastory Pasturk Bistrot Bostrom
شوارزنغر شوارزنيجر شوارتزنجر	→	XFRTSNKR	→	Schwarzenegger
قذافي	→	KTF	→	Qadhafi Gadafi Gaddafi Kadafi Ghaddafi Qaddafi Katif Qatif

Dictionary Expansion

- OOV word is analyzable by BAMA (Buckwalter 2004)
- Add phrase table entries for OOV translating to all inflected forms of the BAMA English gloss
- Assign equal very low probabilities to all entries

الموسيقيون	→	موسيقي	→	musical	→	musical musicals
			→	musician	→	musician musicians
المخطئة	→	مخطئ	→	mistaken	→	mistaken
			→	at fault	→	at fault at faults
جلستم	→	جلس	→	sit	→	sit sits sat sitting

REMOOV Evaluation

- Medium Set
  - 4.1 M words
  - Average token OOV is 2.9%
- All techniques improve on baseline
  - TransEx < MorphEx < DictEx < SpellEx
- Combinations improve on combined techniques
  - Least improving combination (on average): MorphEx+DictEx
  - Most improving combination (on average): DictEx+TransEx
- Combining all improves most

BLEU Scores

	MT03	MT04	MT05
BASELINE	44.20	40.60	42.86
TRANSEX	44.83	40.90	43.25
MORPHEX	44.79	41.18	43.37
DICTEX	44.88	41.24	43.46
SPELLEX	45.09	41.11	43.47
MORPHEX+DICTEX	45.00	41.38	43.54
SPELLEX+dMORPHEX	45.28	41.40	43.64
SPELLEX+TRANSEX	45.43	41.24	43.75
DICTEX+TRANSEX	45.30	41.43	43.72
ALL	**45.60**	**41.56**	**43.95**
Absolute improvement	*1.4*	*0.96*	*1.09*
Relative improvement	*3.17*	*2.36*	*2.54*

REMOOV Evaluation

- Learning Curve Evaluation
  - Different techniques do better under different size conditions
  - Even with 10 times data, OOV handling techniques still help
- Error Analysis
  - Hardest cases are Names
  - 60% of time, OOV handling is acceptable

<u>MT04 BLEU Scores</u>

	1%	10%	100%	1000%
Baseline	13.40	31.07	40.60	42.06
TransEX	13.80	31.78	40.90	42.10
SpellEX	14.02	31.85	41.11	42.25
MorphEX	15.06	32.29	41.18	42.16
DictEx	**20.09**	**33.56**	41.24	42.14
ALL	18.17	33.41	**41.56**	**42.29**
Best Absolute	6.69	2.49	0.96	0.23
Best Relative	49.93	8.01	2.36	0.55

	PN	NOM	V	
Good	26 (40%)	41 (73%)	17 (85%)	*60%*
Bad	39 (60%)	15 (27%)	3 (15%)	*40%*
	46%	*40%*	*14%*	*100%*

OOV Handling Examples

- Foreign name
  - Before: … and president of ecuador lwt$yw gwtyryz .
  - After: … and president of ecuador lucio gutierrez .
- Dual noun
  - Before: … headed the mission to qrytyn in the north .
  - After: … headed the mission to villages in the north .
- Dual verb
  - Before: … baghdad and riyadh , which qTEtA their diplomatic relations …
  - After: … baghdad and riyadh , which sever their diplomatic relations …
- Spelling error
  - Before: … but mHAdtAt between palestinian factions …
  - After: … but talks between palestinian factions …

Road Map

- Machine Translation for Arabic
  - Tokenization for Arabic to English MT
  - OOV Reduction
  - **Dialect to English MT through MSA Pivoting**

Arabic Dialect Machine Translation

- Problems
  - Limited resources
    - Small Dialect-English corpora & no Dialect-MSA corpora
  - Non-standard orthography
  - Morphological complexity
- Solutions
  - Rule-based segmentation (Riesa et al. 2006)
  - Minimally supervised segmentation (Riesa and Yarowsky 2006)
  - Dialect-MSA lexicons (Chiang et al. 2006, Maamouri et al. 2006)
  - Pivoting on MSA (Sawaf 2010, Salloum and Habash, 2011)
    - Elissa 1.0 (Salloum & Habash, 2012)
  - Crowdsourcing Dialect-English corpora (Zbib et al., 2012)

Elissa 1.0

- Dialectal Arabic to MSA MT System
- Output
  - MSA top-1 choice, n-best list or map file
- Components
  - Dialectal morphological analyzer (ADAM) (Salloum and Habash, 2011)
  - Hand-written morphological transfer rules & dictionaries
  - MSA language model
- Evaluation (DA-English MT)
  - MADA preprocessing (ATB scheme)
  - Moses trained for MSA-English MT
  - 64 M words training data
  - Best system only processes MT OOVs and ADAM dialect-only words
  - Top-1 choice of MSA
  - Results in BLEU

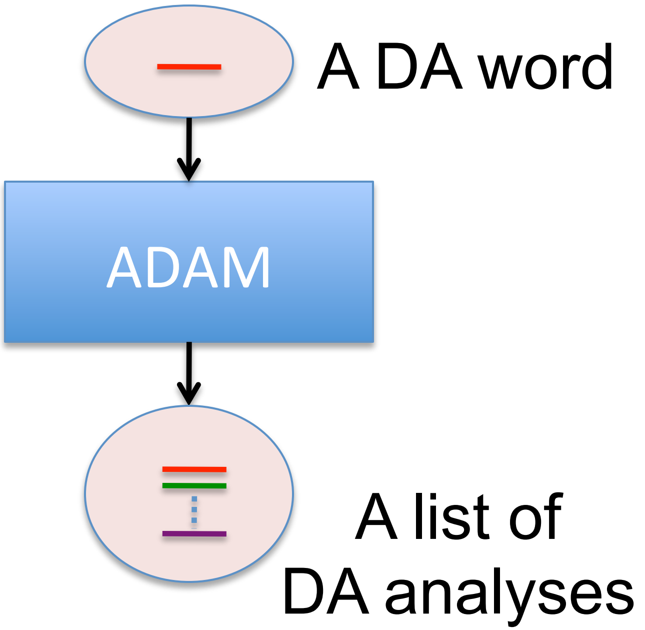
Rule-Based Transfer System

ADAM: Analyzer for Dialectal Arabic Morphology

Output: **lemma** and **feature-value** pairs

Built by **extending BAMA** database adding dialectal affixes and clitics, but no stems.

e.g., sa/FUT → Ha/FUT



Example

<table>
<tr><td></td><td colspan="5">wmAHyktbwlw وماحيكتبولو
“and they will not write to him”</td></tr>
<tr><td rowspan="2">Analysis</td><td>Proclitics</td><td>[Lemma & Features]</td><td>Enclitics</td><td></td><td></td></tr>
<tr><td>w+
conj+
and+ | mA+
neg+
not+ | H+
fut+
will+</td><td>y-ktb-w
[katab IV subj:3MP voice:act]
they write</td><td>+l
+prep
+to</td><td>+w
$+pron_{3MS}$
+him</td><td></td></tr>
<tr><td rowspan="3">Transfer</td><td colspan="2">Word 1</td><td>Word 2</td><td colspan="2">Word 3</td></tr>
<tr><td>Proclitics</td><td>[Lemma& Features]</td><td>[Lemma & Features]</td><td>[Lemma & Features]</td><td>Enclitics</td></tr>
<tr><td>conj+
and+</td><td>[lan]
will not</td><td>[katab IV subj:3MP voice:act]
they write</td><td>[li]
to</td><td>$+pron_{3MS}$
+him</td></tr>
<tr><td rowspan="2">Generation</td><td>w+</td><td>ln</td><td>yktbwA</td><td>l</td><td>+h</td></tr>
<tr><td colspan="5">wln yktbwA lh ولن يكتبوا له</td></tr>
</table>

Example

wmAHyktbwlw وماحيكتبولو

“and they will not write to him”

Analysis

Proclitics			[Lemma & Features]	Enclitics	
w+ conj+ and+	mA+ neg+ not+	H+ fut+ will+	y-ktb-w [katab IV subj:3MP voice:act] they write	+l +prep +to	+w +pron$_{3MS}$ +him

Transfer

Word 1		Word 2	Word 3	
Proclitics	**[Lemma& Features]**	**[Lemma & Features]**	**[Lemma & Features]**	**Enclitics**
conj+ and+	[lan] will not	[katab IV subj:3MP voice:act] they write	[li] to	+pron$_{3MS}$ +him

Generation

w+	ln	yktbwA	l	+h

wln yktbwA lh ولن يكتبوا له

Elissa 1.0: DA to MSA translation

Direct Translation of Dialectal Arabic (DA)	
Dialectal Arabic	بهالحالة ماحيكتبولو شي عحيط صفحتو لأنو ماخبرهن يوم اللي وصل عالبلد
DA-English Human Transaltion	In this case, they will not write on his page wall because he did not tell them the day he arrived to the country.
Arabic-English Google Translate	Bhalhalh Mahiketbolo Shi Ahat Cefhto to Anu Mabrhen day who arrived Aalbuld.

Pivoting on Modern Standard Arabic (MSA) using Elissa	
DA-MSA Elissa Translation	في هذه الحالة لن يكتبوا شي علي حائط صفحته لانه لم يخبرهم يوم الذي وصل الي البلد
Arabic-English Google Translate	In this case it would not write something on the wall yet because he did not tell them the day arrived in the country.

Elissa 1.0

- Dialectal Arabic to MSA MT System
- Output
  - MSA top-1 choice, n-best list or map file
- Components
  - Dialectal morphological analyzer (ADAM) (Salloum and Habash, 2011)
  - Hand-written morphological transfer rules & dictionaries
  - MSA language model
- Evaluation (DA-English MT)
  - MADA preprocessing (ATB scheme)
  - Moses trained for MSA-English MT
  - 64 M words training data
  - Best system only processes MT OOVs and ADAM dialect-only words
  - Top-1 choice of MSA
  - Results in BLEU

System	Dev. Set	Blind Test
Baseline	37.20	38.18
Elissa + Baseline	37.86	38.80

Challenges for Statistical MT

- Data Sparsity
  - Training models need a lot of data
  - Genre and domain sensitive
  - Worse for language with rich morphology
  - Some language pairs have little to no parallel data

Challenges for Statistical MT

- Data Sparsity
  - Training models need a lot of data
  - Genre and domain sensitive
  - Worse for language with rich morphology
  - **Some language pairs have little to no parallel data**
- Solution: Pivoting (aka Bridging)
  - Pivot the translation through a third language
  - Condition: abundance of parallel corpora of the pivot language with the source and target languages
  - Best pivot language today?

ENGLISH

Pivoting in Google Translate

Pivoting in Google Translate

Pivoting in Google Translate

Agreement in Hebrew-English-Arabic Pivot Translation

Hebrew Input	ארבעה דורות, שלוש מורות, חלום אחד: מרים כהן הייתה מחנכת 31 שנה, בתה אילנה מנהלת בית ספר, ונכדתה חווה היא מורה לחינוך גופני. גם הנינה, שרה, בת 6, רוצה להצטרף לעיסוק המשפחתי.
English Pivot	Four generations, three teachers, one dream: Miriam Cohen has been an educator for 31 years, [her] daughter Ilana [is a] school principal, and her granddaughter Eve is a physical education teacher. Also [the] great-granddaughter, Sarah, age 6, wants to join the family occupation.
Arabic Output	أربعة أجيال، وثلاثة مدرسين ، حلم واحد : لقد تم ميريام كوهين معلما لمدة 31 عاما، ابنة إيلانا مديرة المدرسة وحواء حفيدتها هو مدرس التربية البدنية. أيضا أحفاد، سارة ، سن 6 ، يريد ان ينضم الى أسرة الاحتلال. *Four generations, three [male] teachers, one dream: Miriam Cohen [he] finished a [male] educator for 31 years, the daughter of Ilana is the school principal, and her granddaughter Eve, [he] is the physical education teacher. Also great-grandchildren Sarah, age 6, [he] wants that [he] joins the family of [military] occupation.*

- Most lexical translation is correct
- Half of the errors in Arabic involve gender
- A third of the errors in Arabic involve the determiner Al+
- Source of ambiguity is mostly English, and some Hebrew

Pivoting Strategies

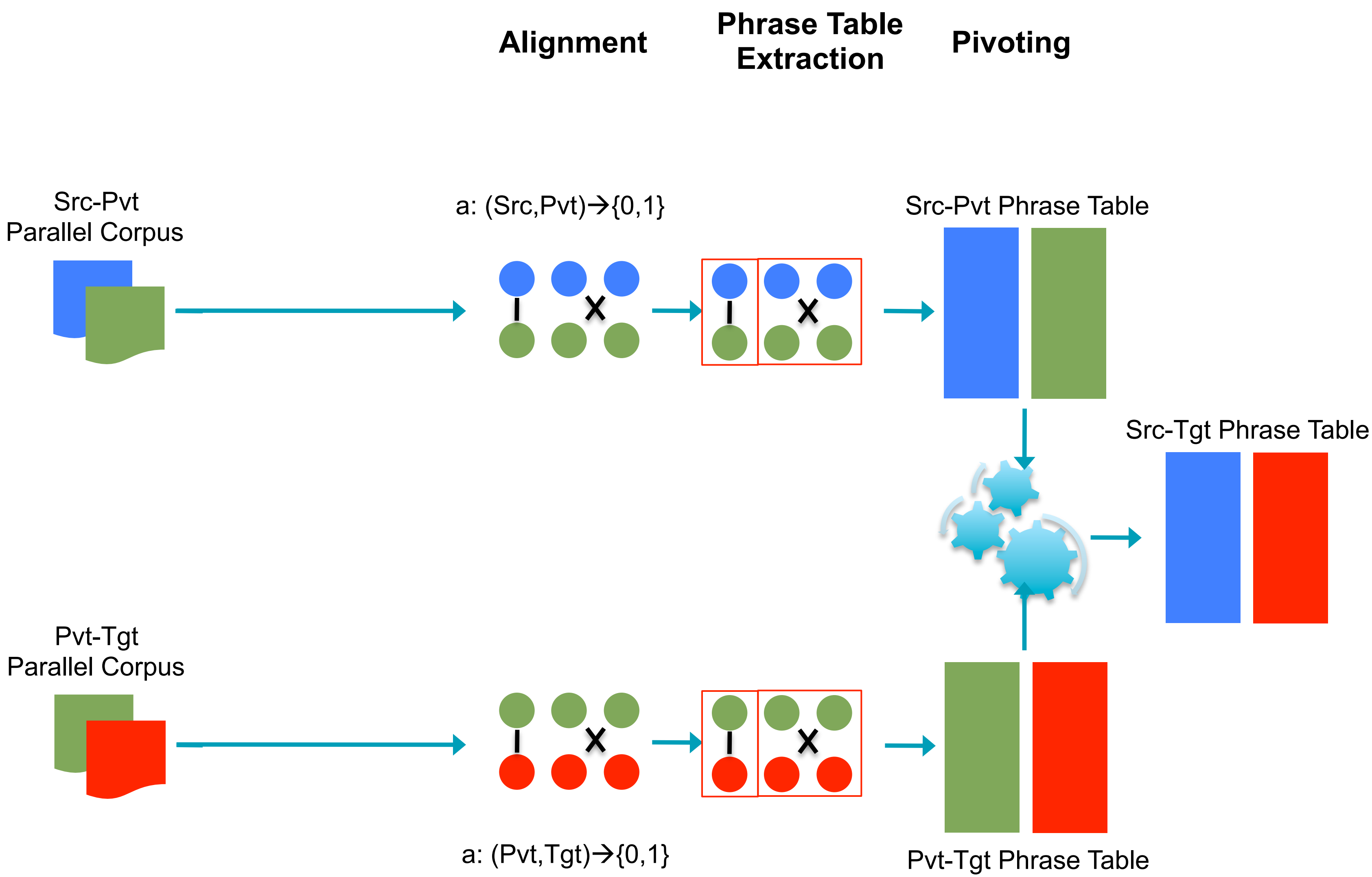
- ***Sentence Pivoting***

  Translate source sentence to pivot, and then from pivot to target

- ***Phrase Pivoting*** (best performing technique)

  Induce a source-to-target translation model from source-to-pivot and pivot-to-target translation models

Phrase Pivoting



Language Comparison

	English	Persian	Arabic
Family	Indo-European	Indo-European	Semitic
Script	Roman	Arabic	Arabic
Word Order	SVO	SOV	SVO/VSO
Morphology	Poor	Rich	Rich

- Every pair of languages share some aspects and differ in others
- Pivoting between two morphologically rich languages is challenging as information drops when transferred through English

Tokenization

Language	Best Tokenization	System	Baseline Bleu	Tokenized Bleu
Arabic	Penn Arabic Treebank	En→ Ar (60 M) [El Kholy and Habash, 2012]	31.30	32.24 (+1.0)
Persian	VerbStem	Pr→En (4M) [Rasooli et al., 2013]	31.40	33.30 (+2.0)

وسيكتبونها wsyktbwnhA
و+ س+ يكتبون +ها w+ s+ yktbwn +hA

and they will write it

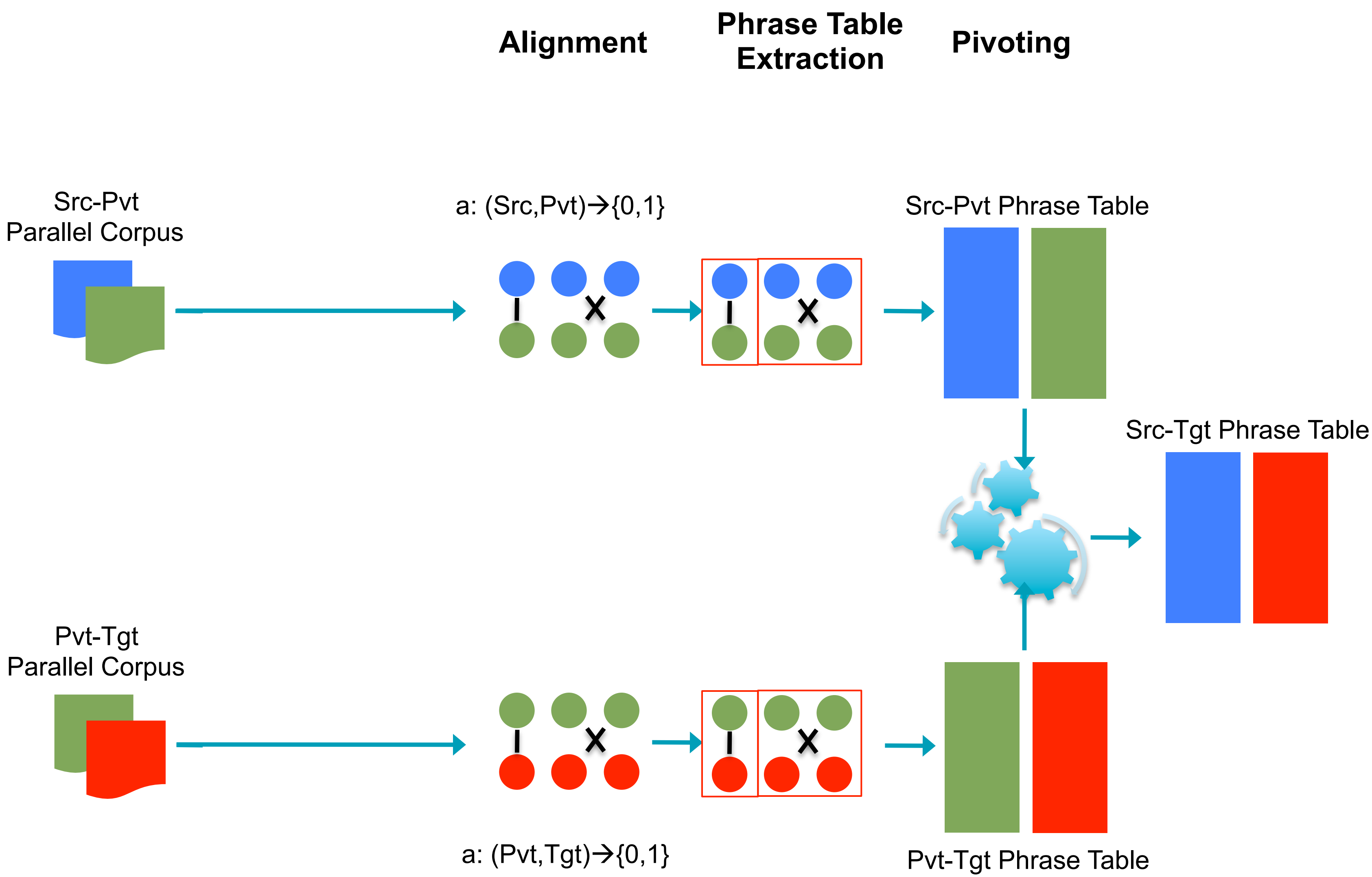
Tokenization

Language	Best Tokenization	System	Baseline Bleu	Tokenized Bleu
Arabic	Penn Arabic Treebank	En→ Ar (60 M) [El Kholy and Habash, 2012]	31.30	32.24 (+1.0)
Persian	VerbStem	Pr→En (4M) [Rasooli et al., 2013]	31.40	33.30 (+2.0)
Hebrew	HTAG	He→En (1M) [Nimesh and Habash, 2012]	19.31	22.79 (+3.5)

Bleu

- Precision-based metric
- 0-100%
- Higher is better
- De facto standard in the field

Phrase Pivoting



Language Independent Connectivity Strength Features

Persian: AyjAd cnd šrkt mštrk ‘ایجاد چند شرکت مشترک’

‘Establish few joint companies’

English: joint ventures

Arabic: bçD šrkAt AlmqAwlAt fy Albld ‘بعض شركات المقاولات في البلد’

‘Some construction companies in the country’

Persian: AçtmAd myAn dw kšwr ‘اعتماد میان دو کشور’

‘trust between the two countries’

English: trust between the two countries

Arabic: Alθqħ byn Aldwltyn ‘الثقة بين الدولتين’

‘the trust between the two countries’

Language Independent Connectivity Strength Features

We add ***two new language-independent features*** to the ***log linear space*** of features in order to model the ***quality*** of the pivot phrase pairs.

- ***Connectivity Strength Features***
- Source Connectivity Strength (SCS)
- Target Connectivity Strength (TCS)

$$SCS = \frac{|A|}{|S|} \qquad TCS = \frac{|A|}{|T|}$$

S is the set of source words in a given phrase pair in the pivot phrase table
T is the set of the equivalent target words
A is the word alignment between ***S*** and ***T***

Language Independent Connectivity Strength Features

Persian: AyjAd cnd šrkt mštrk ‘ایجاد چند شرکت مشترک’

‘Establish few joint companies’

English: joint ventures

SCS=0.25 and TCS=0.2

Arabic: bçD šrkAt AlmqAwlAt fy Albld ‘بعض شركات المقاولات في البلد’

‘Some construction companies in the country’

Persian: AçtmAd myAn dw kšwr ‘اعتماد میان دو کشور’

‘trust between the two countries’

English: trust between the two countries

SCS=1.0 and TCS=1.0

Arabic: Alθqħ byn Aldwltyn ‘الثقة بين الدولتين’

‘the trust between the two countries’

Results (Pr-En-Ar)

Parallel Data: Pr-En: 4M words; En-Ar: 60M words

Evaluation Set: 536 Pr-Ar sentences with 3-references

Results: connectivity strength features boost the performance

Pivot Scheme	BLEU
Phrase Pivoting	20.5
Phrase Pivoting + Connectivity	21.1

BLEU score is statistically significant over the baseline

Next Time

- J+M Chap 19
- Come with questions about the MT assignment