

Better Arabic Parsing

Baselines, Evaluations, and Analysis

Spence Green and Christopher D. Manning

Stanford University

August 27, 2010



Common Multilingual Parsing Questions...

Is language X “**harder**” to parse than language Y?

Common Multilingual Parsing Questions...

Is language X “**harder**” to parse than language Y?

- ▶ Morphologically-rich X

Common Multilingual Parsing Questions...

Is language X “**harder**” to parse than language Y?

- ▶ Morphologically-rich X

Is treebank X “**better/worse**” than treebank Y?

Common Multilingual Parsing Questions...

Is language X “**harder**” to parse than language Y?

- ▶ Morphologically-rich X

Is treebank X “**better/worse**” than treebank Y?

Does feature Z “**help more**” for language X than Y?

Common Multilingual Parsing Questions...

Is language X “**harder**” to parse than language Y?

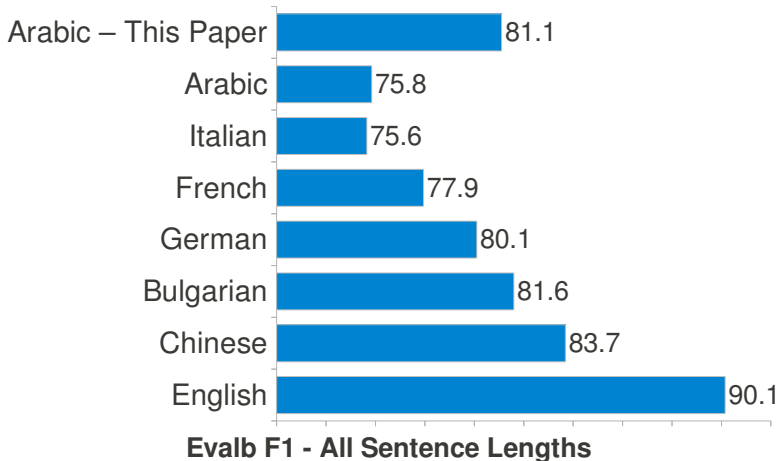
- ▶ Morphologically-rich X

Is treebank X “**better/worse**” than treebank Y?

Does feature Z “**help more**” for language X than Y?

- ▶ Lexicalization
- ▶ Morphological annotations
- ▶ Markovization
- ▶ etc.

“Underperformance” Relative to English



(Petrov, 2009)

Why Arabic / Penn Arabic Treebank (ATB)?

Annotation style similar to PTB

Relatively little segmentation (cf. Chinese)

Richer morphology (cf. English)

More syntactic ambiguity (unvocalized)

ATB Details

Parts 1–3 (not including part 3, v3.2)

Newswire only

- ▶ Agence France Presse, Al-Hayat, Al-Nahar

Corpus/experimental characteristics

- ▶ 23k trees
- ▶ 740k tokens
- ▶ Shortened “Bies” POS tags
- ▶ Split: 2005 JHU workshop

Arabic Preliminaries

Diglossia: “Arabic” \rightarrow MSA

Arabic Preliminaries

Diglossia: “Arabic” \rightarrow MSA

Typology: VSO — VOS, SVO, VO also possible

Arabic Preliminaries

Diglossia: “Arabic” → MSA

Typology: VSO — VOS, SVO, VO also possible

Devocalization

فَقَدْ مَنَعَهُمُ الْحَقَّ

Arabic Preliminaries

Diglossia: “Arabic” → MSA

Typology: VSO — VOS, SVO, VO also possible

Devocalization

فَقَدْ مَنَحَهُمُ الْحَقَّ

فقد منحهم الحق

Arabic Preliminaries

Diglossia: “Arabic” → MSA

Typology: VSO — VOS, SVO, VO also possible

Devocalization

فَقَدْ مَنَحَهُمُ الْحَقَّ
فقد منحهم الحق

Segmentation: an analyst's choice!

- ▶ ATB uses clitic segmentation

Arabic Preliminaries

Diglossia: “Arabic” → MSA

Typology: VSO — VOS, SVO, VO also possible

Devocalization

فَقَدْ مَنَحَهُمُ الْحَقَّ
فقد منحهم الحق

Segmentation: an analyst's choice!

- ▶ ATB uses clitic segmentation

ف + قد منح + هم الحق

Syntactic Ambiguity in Arabic

This talk:

- ▶ Devocalization
- ▶ Discourse level coordination ambiguity

Many other types:

- ▶ Adjectives / Adjective phrases
- ▶ Process nominals — *maSdar*
- ▶ Attachment in annexation constructs (Gabbard and Kulick, 2008)

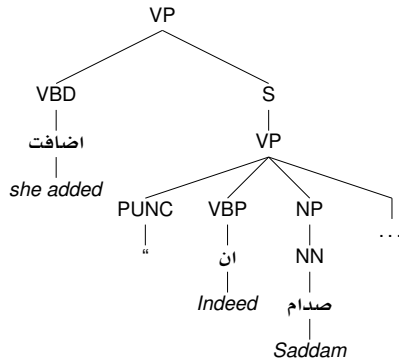
Devocalization: *Inna and her Sisters*

			POS	Head Of
U ¹³ _e I	<i>inna</i>	“indeed”	VBP	VP
U ¹³ _e I	<i>anna</i>	“that”	IN	SBAR
U ^o _e I	<i>in</i>	“if”	IN	SBAR
U ^o _e I	<i>an</i>	“to”	IN	SBAR

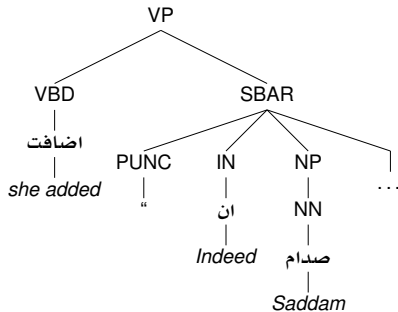
Devocalization: *Inna and her Sisters*

			POS	Head Of
ان	<i>inna</i>	“indeed”	VBP	VP
ان	<i>anna</i>	“that”	IN	SBAR
ان	<i>in</i>	“if”	IN	SBAR
ان	<i>an</i>	“to”	IN	SBAR

Devocalization: *Inna and her Sisters*

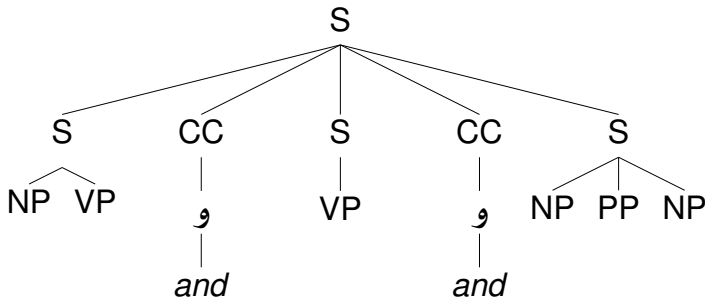


Reference



Stanford

Discourse-level Coordination Ambiguity



- ▶ S < S in 27.0% of dev set trees
- ▶ NP < CC in 38.7% of dev set trees

Discourse-level Coordination Ambiguity

Leaf Ancestor metric reference chains (Berkeley)

- ▶ score $\in [0, 1]$

Score	# Gold	
0.696	34	S < S < VP < NP < PRP
0.756	170	S < VP < NP < CC
0.768	31	S < S < VP < S < VP < PP < IN
0.796	86	S < S < VP < SBAR < IN
0.804	52	S < S < NP < NN

Treebank Comparison

Compared ATB gross corpus statistics to:

- ▶ Chinese — CTB6
- ▶ English — WSJ sect. 2-23
- ▶ German — Negra

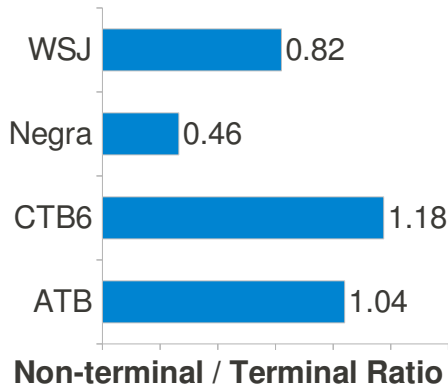
Treebank Comparison

Compared ATB gross corpus statistics to:

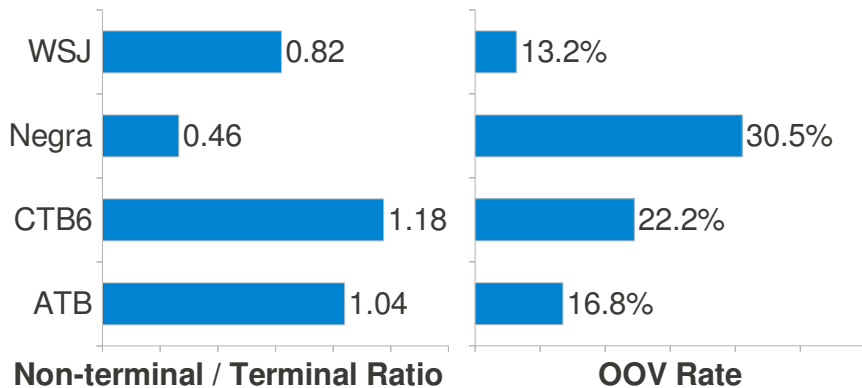
- ▶ Chinese — CTB6
- ▶ English — WSJ sect. 2-23
- ▶ German — Negra

The ATB isn't that unusual!

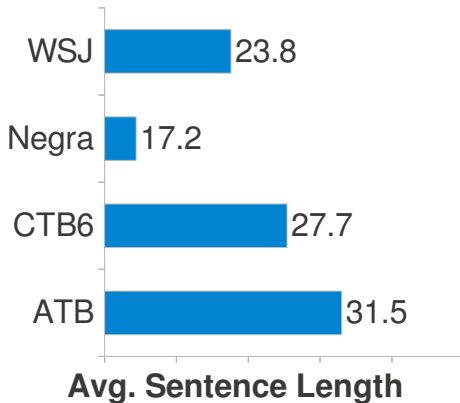
Corpus Features in Favor of the ATB



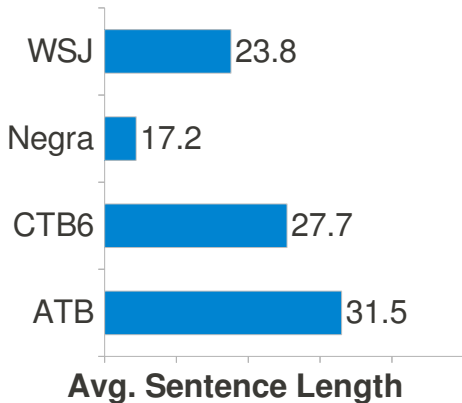
Corpus Features in Favor of the ATB



Sentence Length Negatively Affects Parsing



Sentence Length Negatively Affects Parsing



40 words is not a sufficient limit for evaluation!

Developing a Manually Annotated Grammar

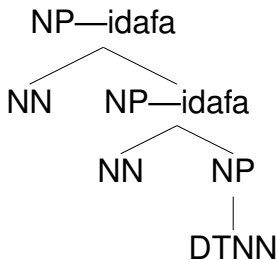
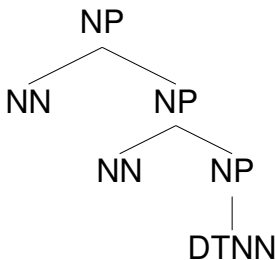
Klein and Manning (2003)–style state splits

- ▶ Human–interpretable
- ▶ Features can inform treebank revision

Developing a Manually Annotated Grammar

Klein and Manning (2003)–style state splits

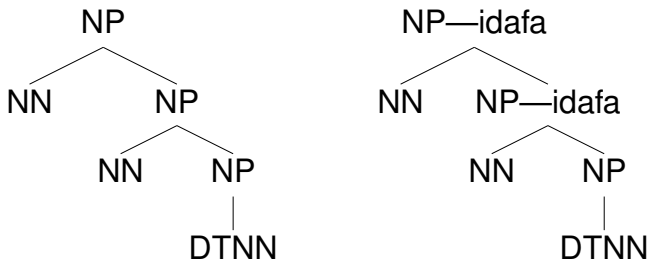
- ▶ Human–interpretable
- ▶ Features can inform treebank revision



Developing a Manually Annotated Grammar

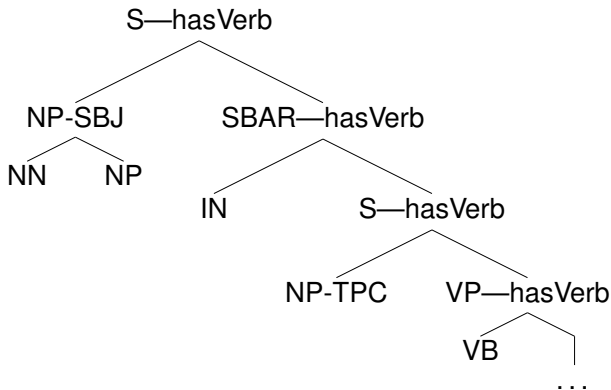
Klein and Manning (2003)–style state splits

- ▶ Human–interpretable
- ▶ Features can inform treebank revision

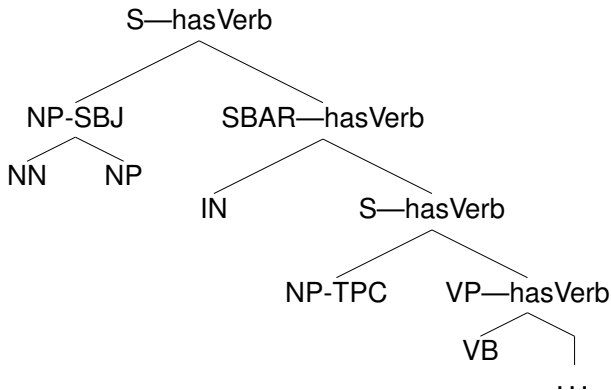


Alternative: automatic splits (Berkeley parser)

Feature: markContainsVerb

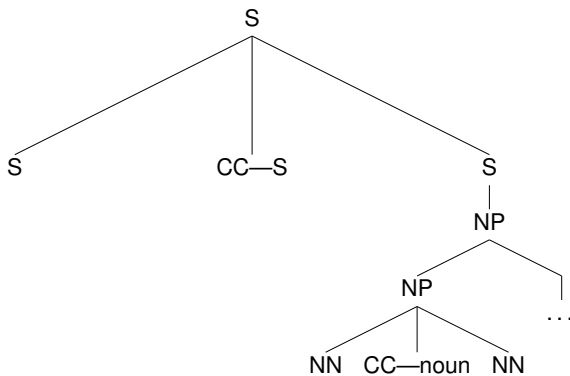


Feature: markContainsVerb

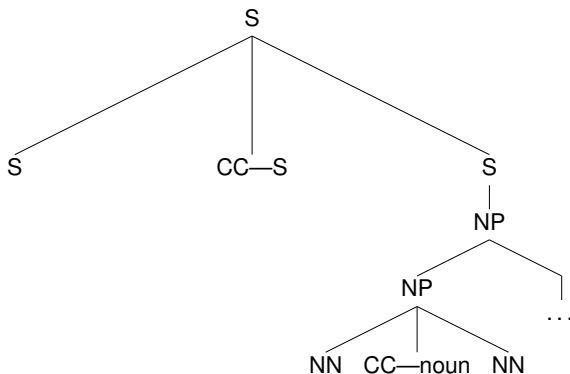


- ▶ +1.18 dev set improvement
- ▶ 16.1% of dev set trees lack verbs

Feature: splitCC



Feature: splitCC



- ▶ +0.21 dev set improvement
- ▶ POS equivalence classes for verb, noun, adjective

Gold Experimental Setup

Evaluation on lengths ≤ 70

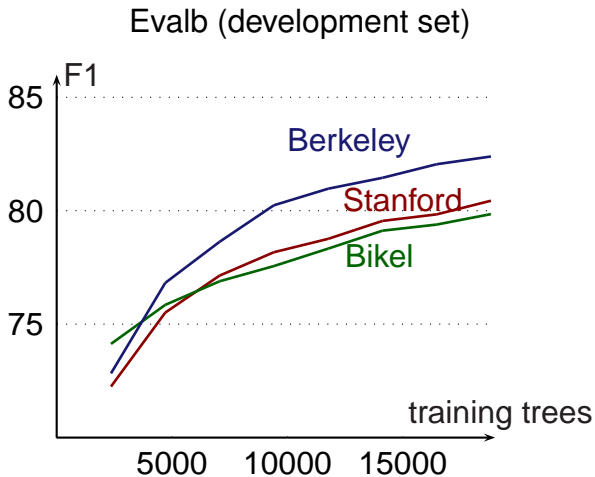
Pre-processing makes a *huge* difference

- ▶ Maintained non-terminal / terminal ratio vv. prior work

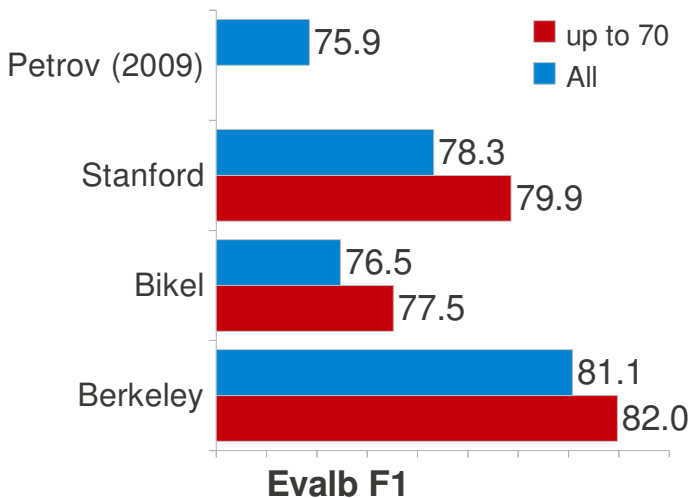
Models:

- ▶ Berkeley
- ▶ Bikel — with pre-tagged input
- ▶ Stanford — with the manual grammar

Learning Curves



Model Comparison



Raw Text Experimental Setup

Pipeline: MADA + Stanford parser

Raw Text Experimental Setup

Pipeline: MADA + Stanford parser

Lattice parsing – effective for Hebrew (Goldberg and Tsarfaty, 2008)

Raw Text Experimental Setup

Pipeline: MADA + Stanford parser

Lattice parsing – effective for Hebrew (Goldberg and Tsarfaty, 2008)

Evaluation on gold (segmented) lengths ≤ 70

Raw Text Experimental Setup

Pipeline: MADA + Stanford parser

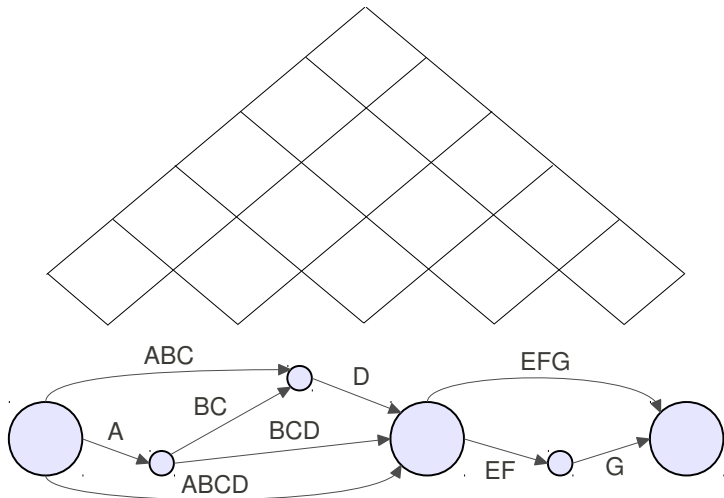
Lattice parsing – effective for Hebrew (Goldberg and Tsarfaty, 2008)

Evaluation on gold (segmented) lengths ≤ 70

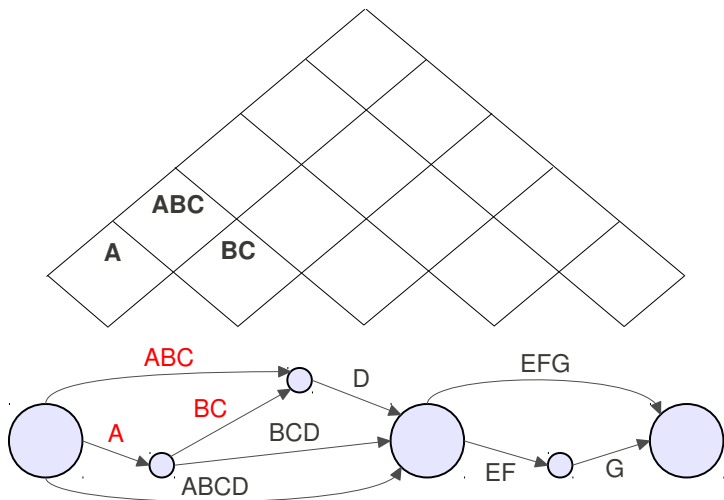
Metric: Evalb without whitespace

- ▶ Requires exact character yield (Tsarfaty, 2006)

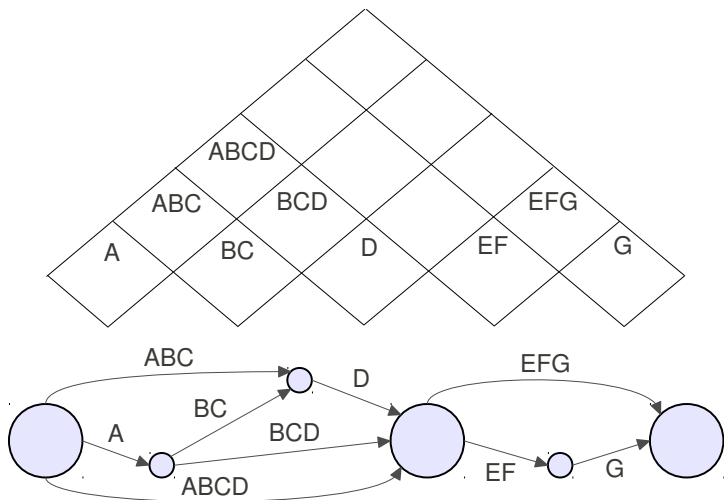
Lattice Parsing — “ABCD EFG”



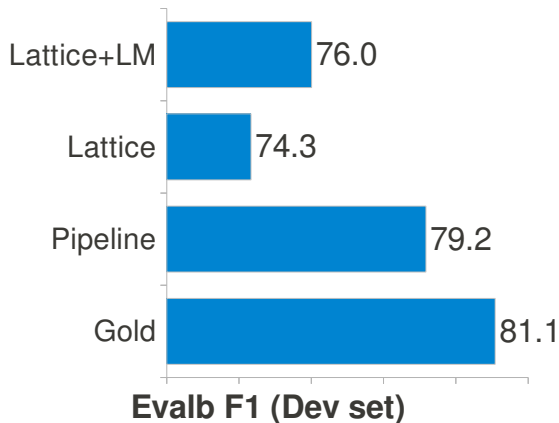
Lattice Parsing — “ABCD EFG”



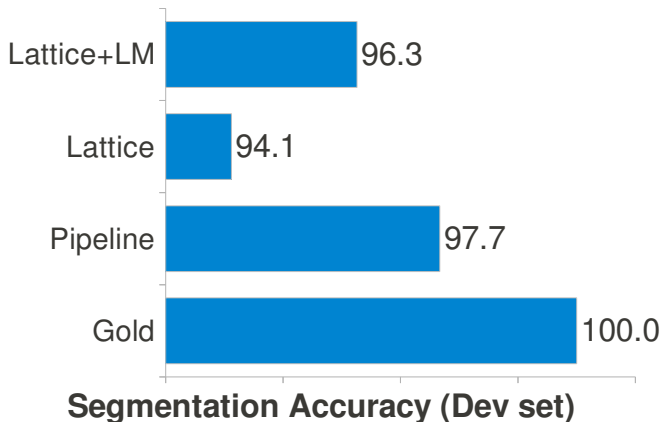
Lattice Parsing — “ABCD EFG”



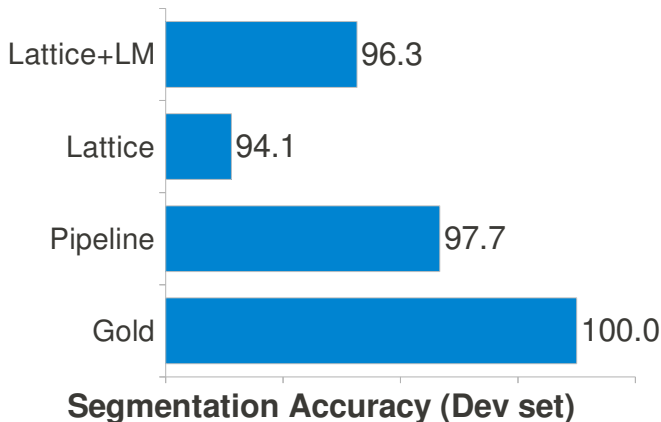
Parsing and Segmentation Results



Parsing and Segmentation Results

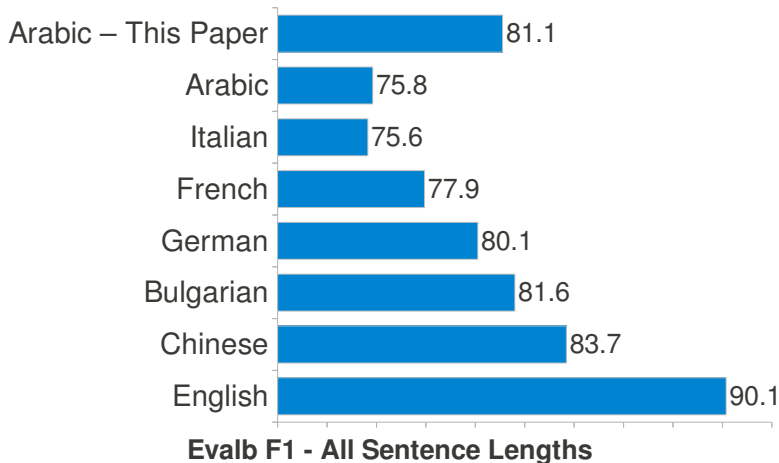


Parsing and Segmentation Results



Comparable segmentation without the effort!

Conclusion

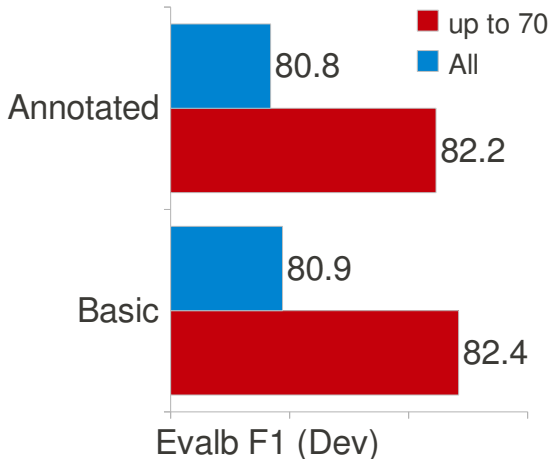


ولكم جزيل الشكر

Thanks.

<http://nlp.stanford.edu/projects/arabic.shtml>

Mixture of Manual and Automatic Annotation



Frequency Matched Strata

	Arabic		English	
	μ freq.	% nuclei	μ freq.	% nuclei
2	2.00	46%	2.00	34%
[3, 4]	3.37	26%	3.37	27%
[5, 9]	6.43	16%	6.49	20%
[10, 49]	18.6	10%	19.0	16%
[50, 500]	110.3	2%	113	3%

Annotation Consistency Evaluation (Again!)

	Nuclei per tree	Sample n-grams	Error %	
			Type	n-gram
WSJ 2-21	0.565	750	16.0%	4.10%
ATB	0.830	658	26.0%	4.10%

95% confidence intervals (type-level):

- ▶ Arabic: [17.4%, 34.6%]
- ▶ English: [8.79%, 23.3%]