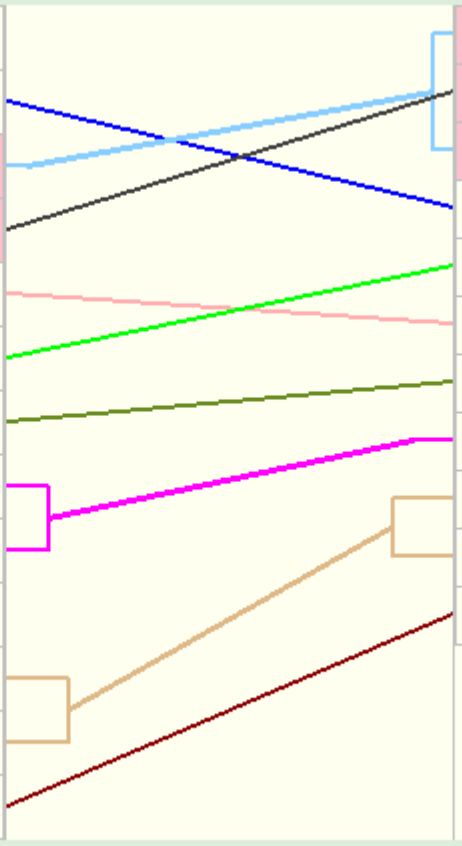
IBM Research

Arabic Parsing for MT

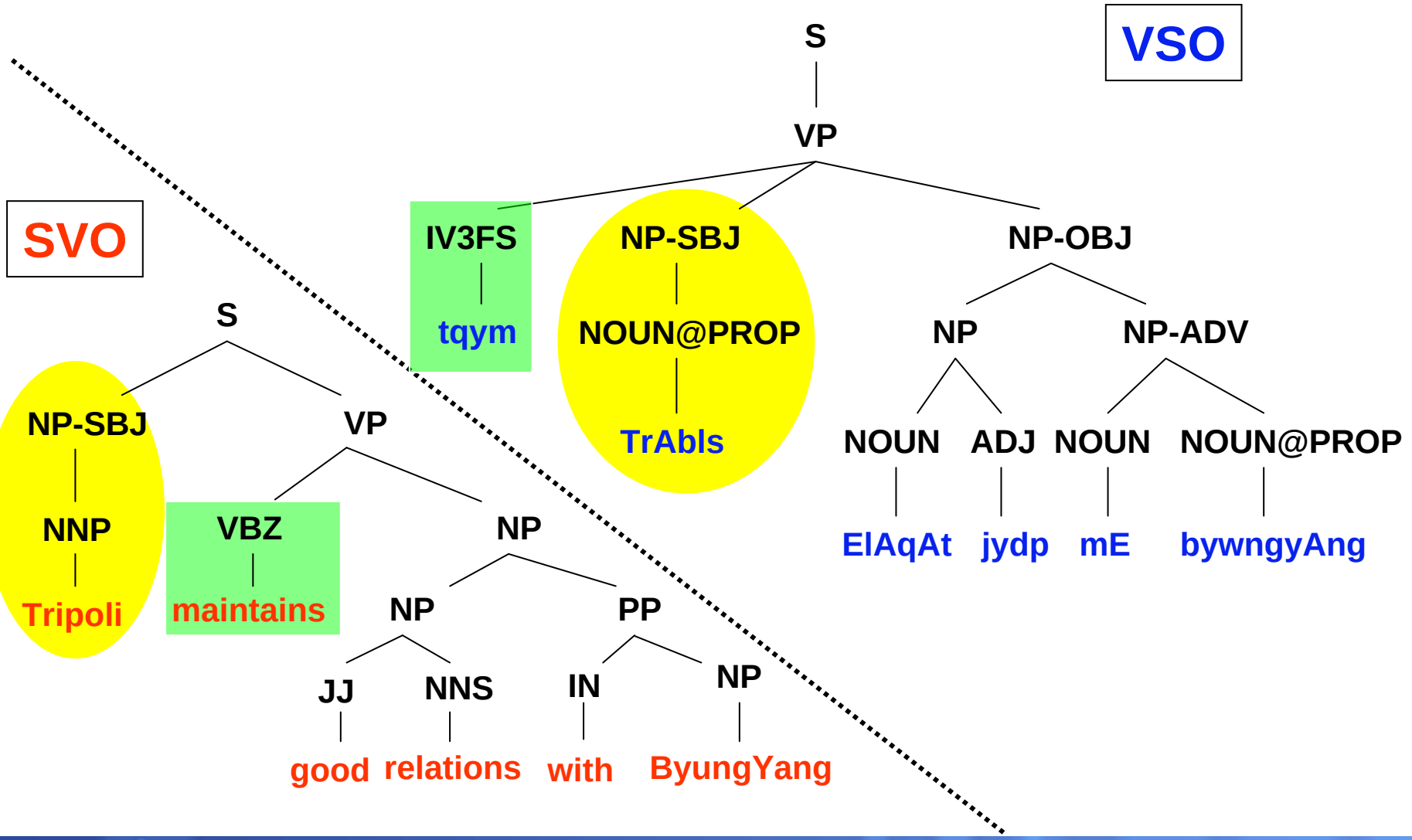
Young-Suk Lee
Abe Ittycheriah

S——CONJ	1	extra	w#
├VP——PV	2	g	ftHt
│ ├NP_SBJ——NOUN	3	g	mrAkz
│ │ └NP——DET_NOUN	4	g	AlAqtrAE
│ ├NP_OBJ——NOUN	5	g	AbwAb
│ │ └NP——POSS@PRON	6	g	+hA
│ └NP_ADV——NOUN	7	g	End
│ └NP–NP–NP–DET_NOUN	8	g	AlsAEp
│ │ └NP——NUM	9	g	$num_(00,06)
│ └PP——PREP	10	extra	b#
│ └NP——DET_NOUN	11	g	Altwqyt
│ └——DET_ADJ	12	g	AlmHly
└——PUNC	13	g	.

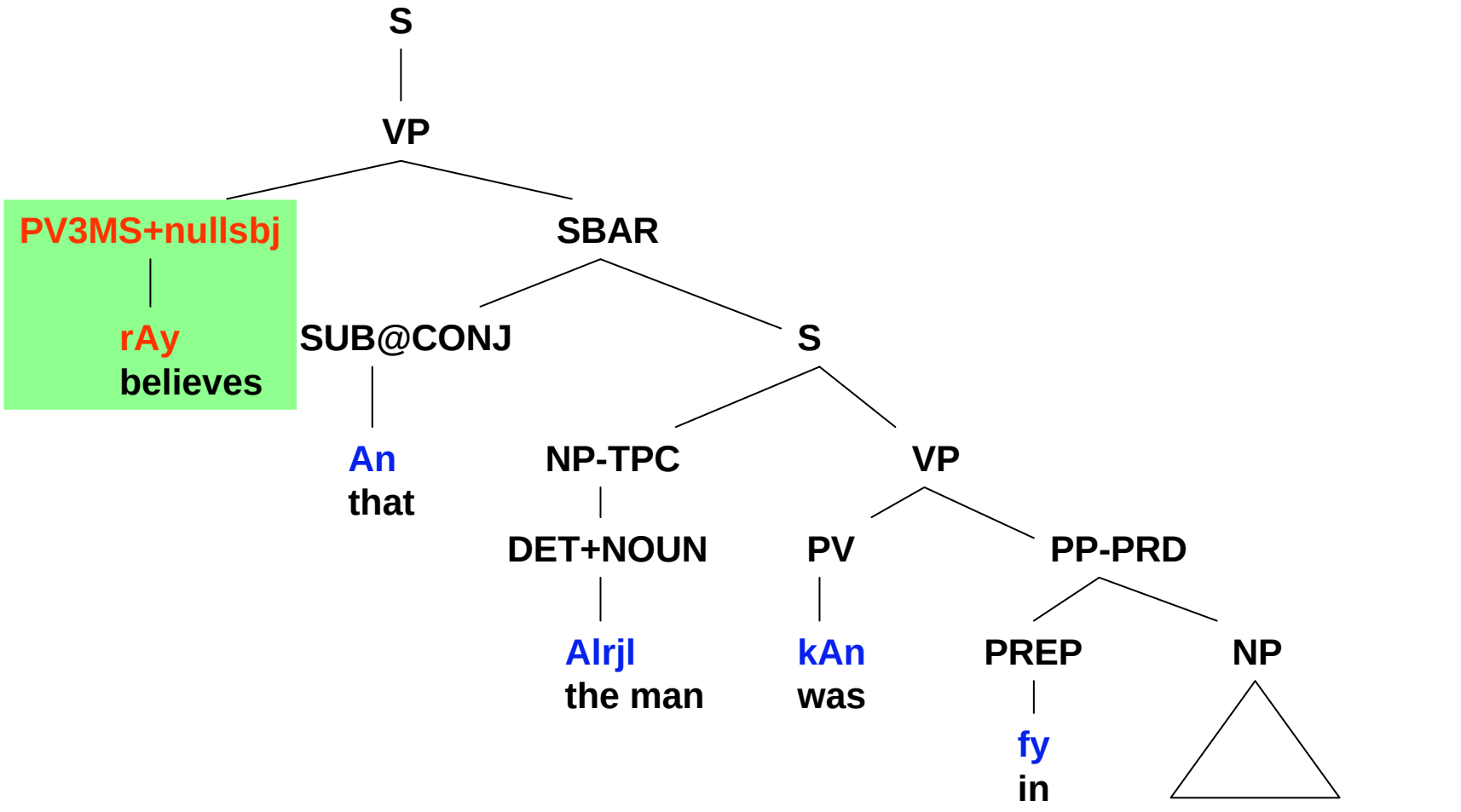


the	g	1
voting	g	2
centers	g	3
opened	g	4
their	g	5
doors	g	6
at	g	7
00,06	g	8
local	g	9
time	g	10
.	g	11

Verb-Subject-Object vs. SVO Word Order



Null Subjects in Arabic



(He) believes that the man was in the wrong place

Arabic Parsing: Experimental Setup

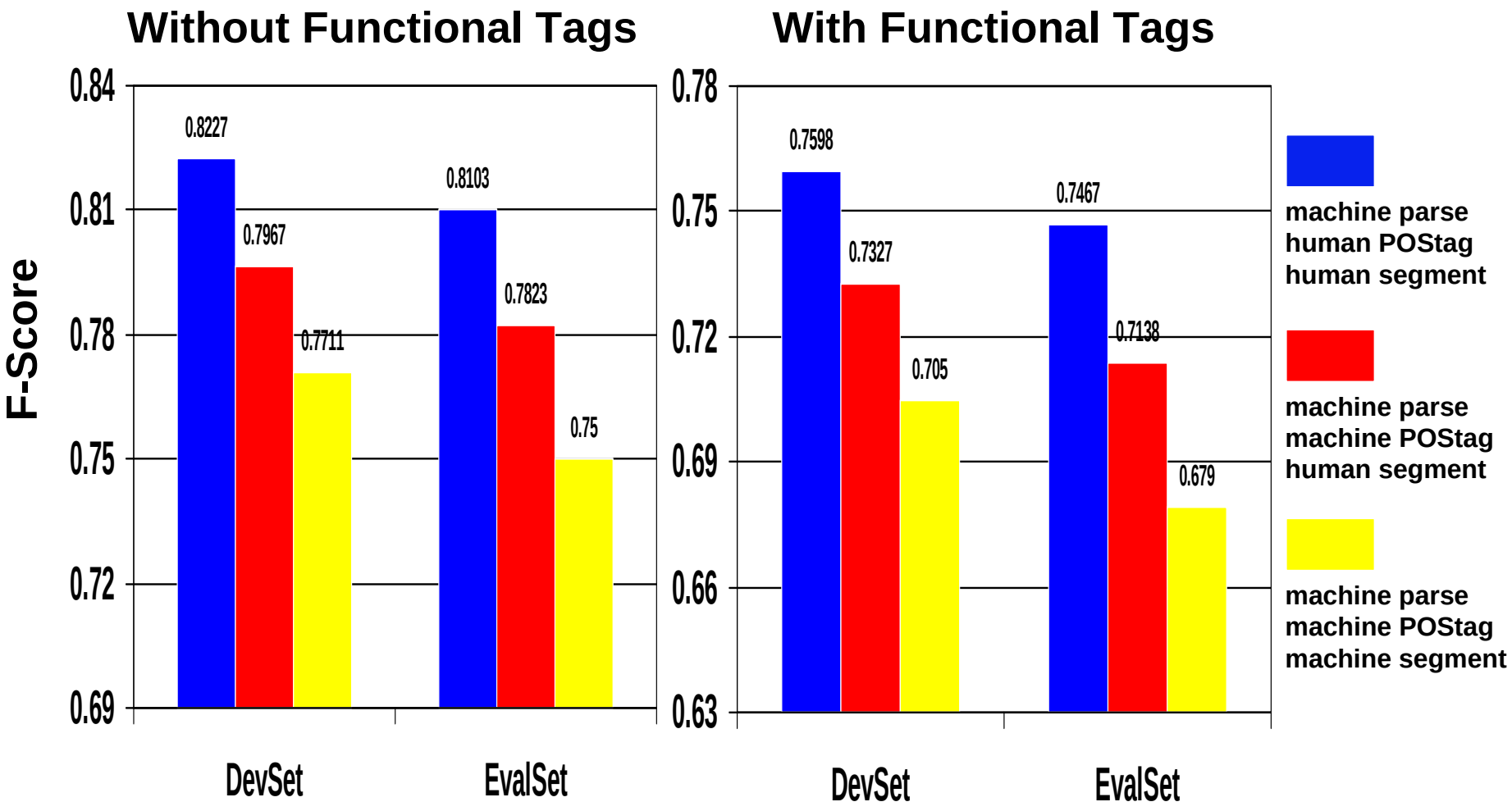
Data Partition

DataSet	Morpheme Count	Source	Sections
Train	658,621 (31 morphs/sent)	ATB1-v4.0 (LDC2008E61) ATB2-v3.0 (LDC2008E62) ATB3-v3.1 (LDC2008E22)	
Development	30,375 (29 morphs/sent)	ATB3-v3.1 (LDC2008E22)	ANN20021015.0101 – ANN20021115.0066
Evaluation	29,048 (29 morphs/sent)	ATB3-v3.1 (LDC2008E22)	ANN20021115.0068 – ANN20021215.0045

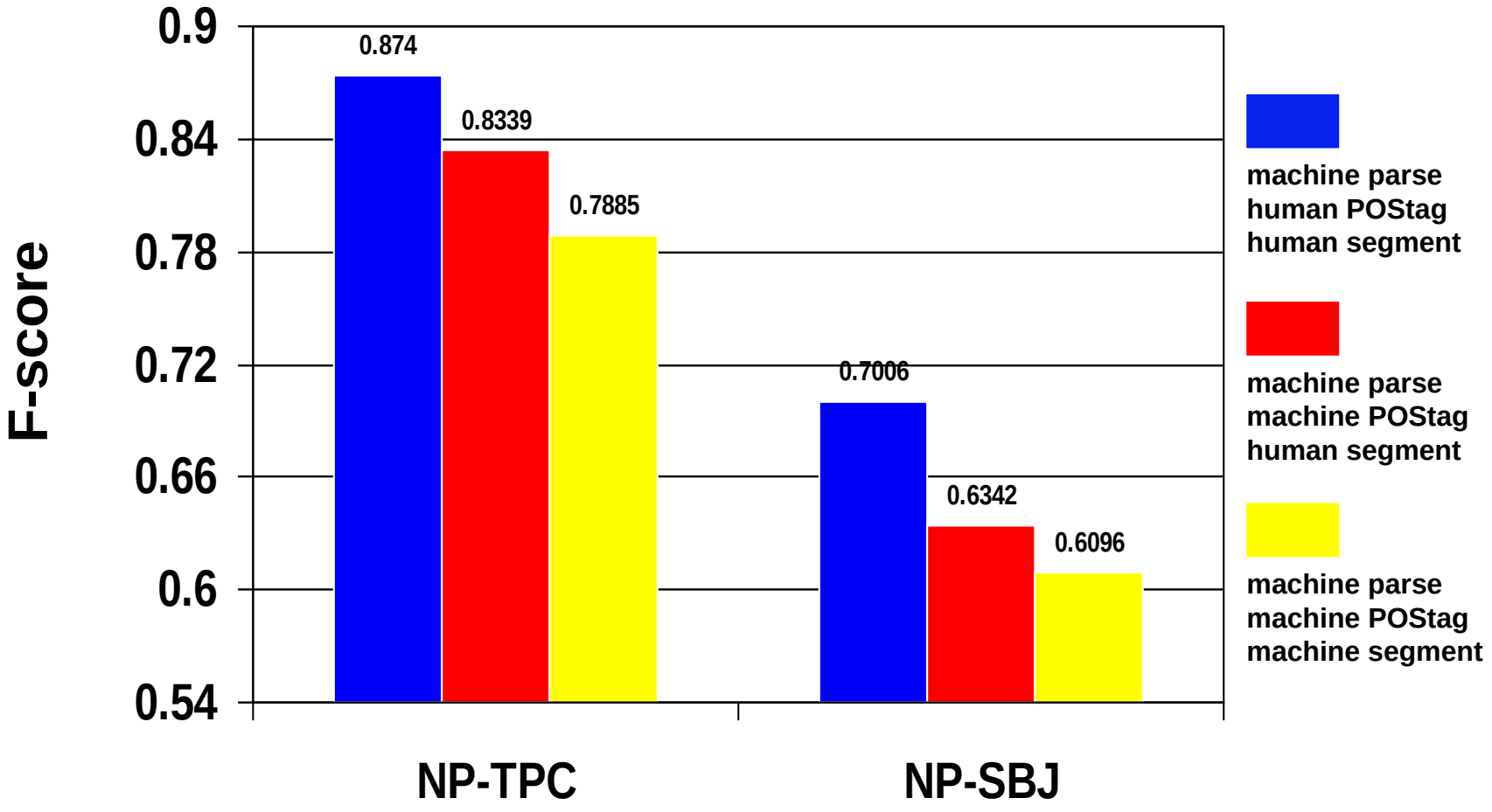
Approach

- Build on the state-of-the-art maximum entropy parser
- Focus on the features critical for SMT performance improvement
  - Accurately parse subject/topic noun phrase
  - Identify null subjects and their properties, e.g. number, gender

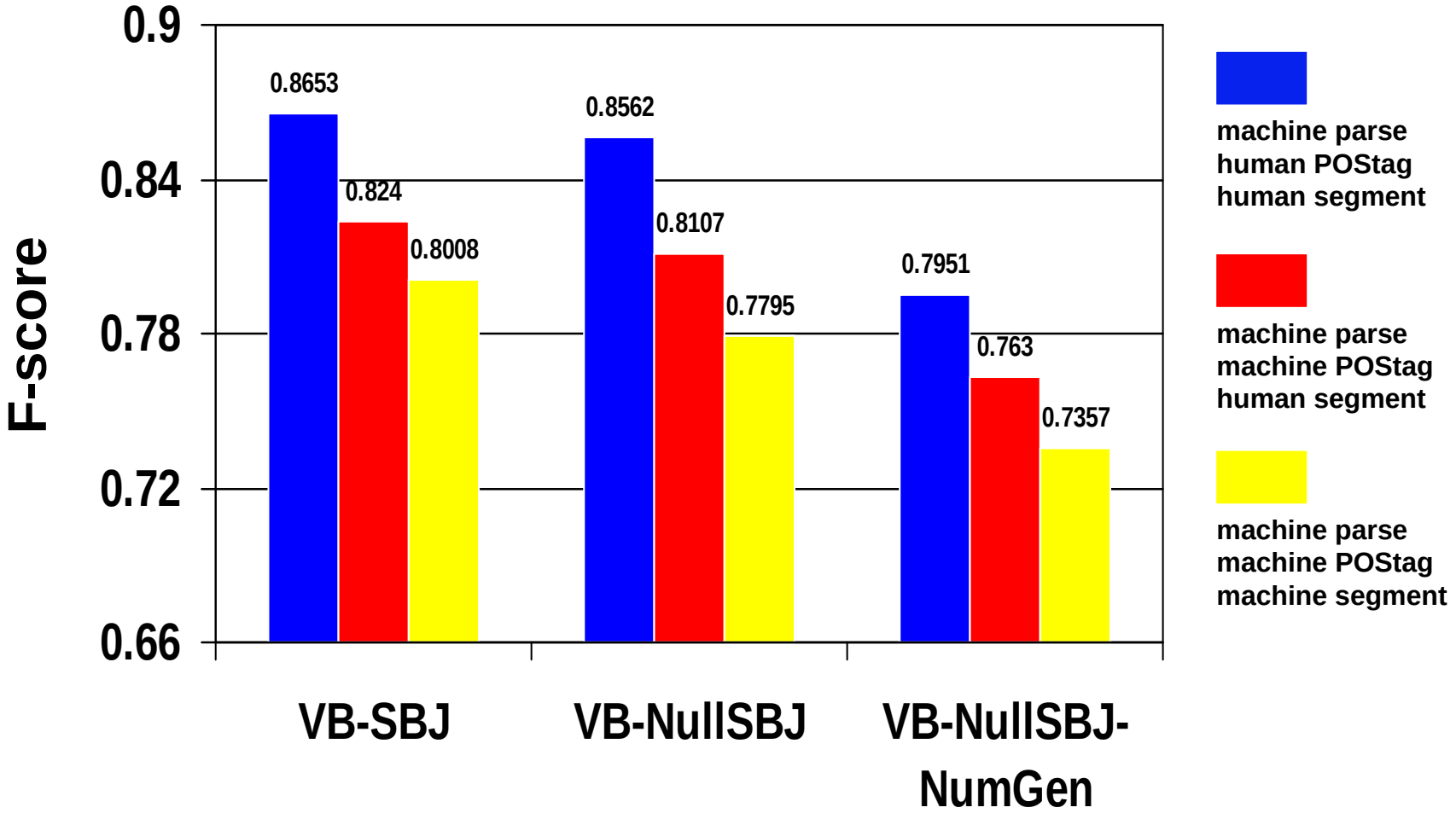
Parser Performance



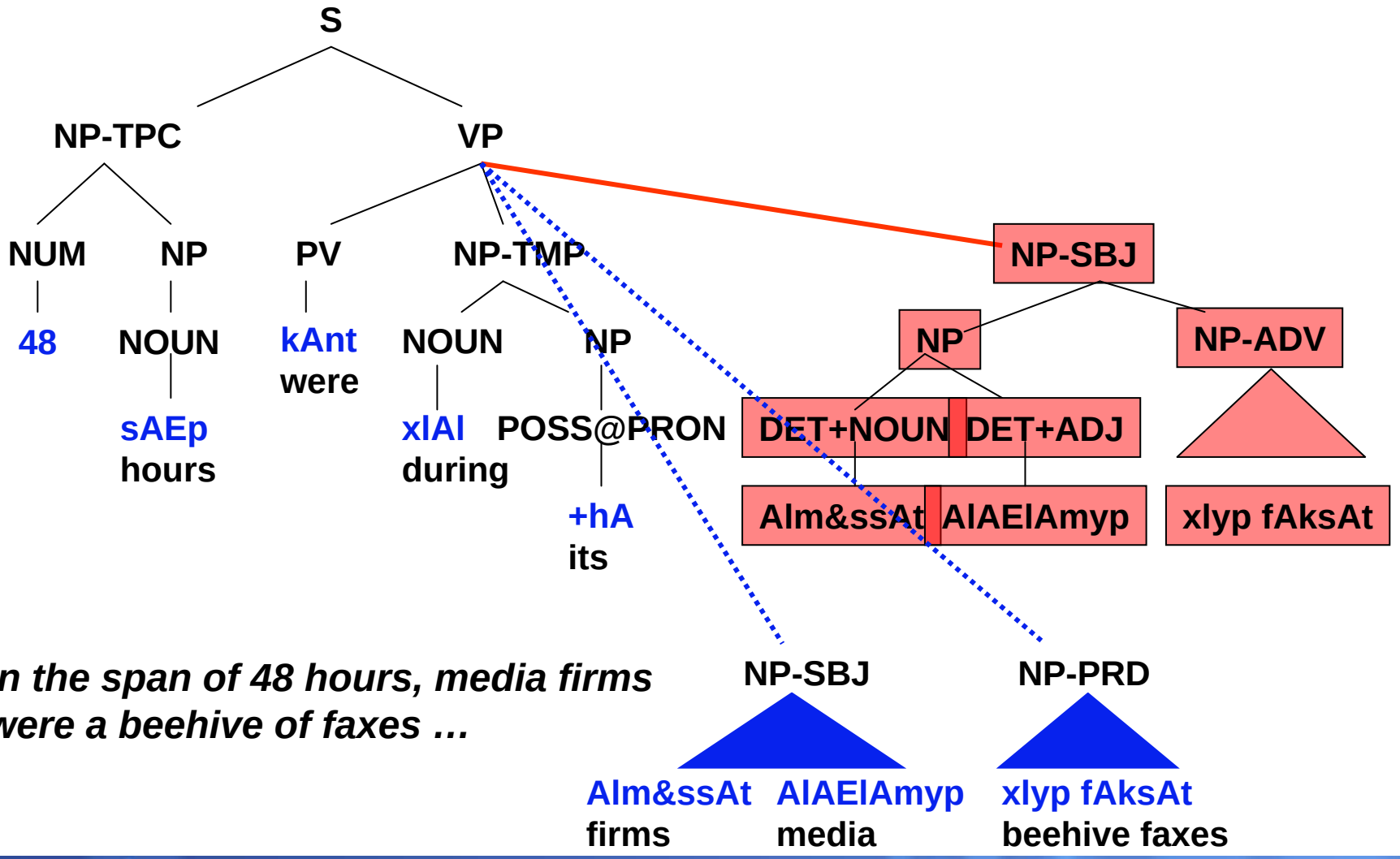
NP-SBJ/TPC Parsing Accuracy: DevSet



Null Subject Recovery Accuracy: DevSet

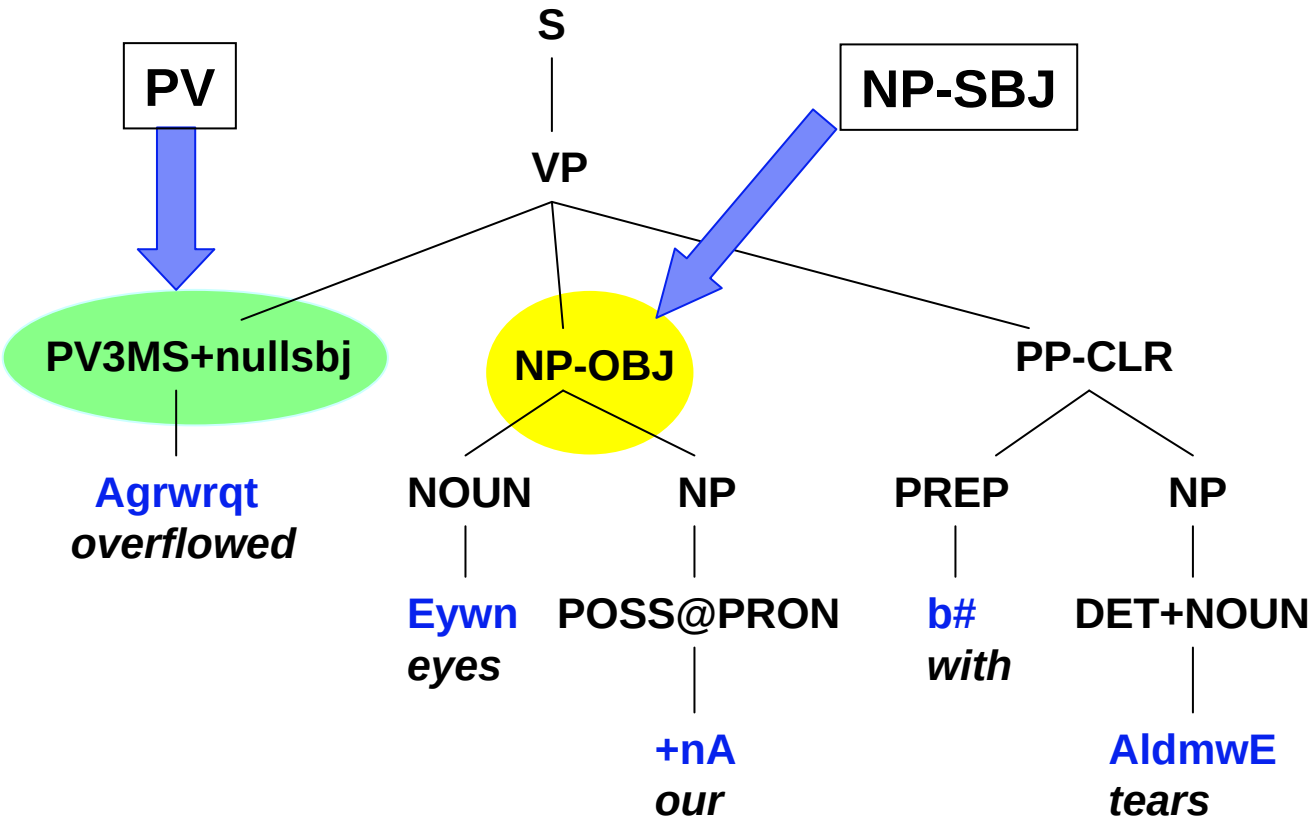


Parsing Error: Long Subject Span



In the span of 48 hours, media firms were a beehive of faxes …

Parsing Error: Incorrect VERB-NullSBJ



Our eyes overflowed with tears.

IBM DTM – DEV07 NW

Feature Types	# of feats	TER	Bleu	(T-B)/2
Baseline Phrase Decoder		41.53	52.90	-5.68
Lexical Features	524528	40.89	52.63	-5.87
+Lexical Context Features	2312573	40.09	54.38	-7.15
+Lexical Trigram Context	3402358	40.12	54.28	-7.08
+Segmentation Features	3574173	40.01	54.33	-7.16
+Variable Features	3584318	40.11	54.20	-7.04
+Coverage Features	3719349	40.06	54.30	-7.12
+POS Features	3818413	39.96	55.06	-7.55
+ParseFeats	4063782	39.78	55.37	-7.79