Lexical Syntax for Statistical Machine Translation

Hany Hassan

DCU & IBM

In collaboration with:
Andy Way and Khalil Sima'an

Outline

Outline

Can linguistic syntax improve PBSMT?

(Koehn et al 2003) tried to impose syntactic constituents on phrase extraction

Hierarchical Phrase structure (Chiang 2005)

- Allows for hierarchical phrases
- Handles a range of reordering problems
- The syntax induced is not linguistically motivated.

Syntactified target phrases (Marcu et. al. 2006)

- Induces millions of xRs rules from parallel corpus
- Mismatch between constituent (xRs) and phrase
- Subtrees for phrases: leads to spurious ambiguity in phrase table

Do subtrees/constituents fit well with phrases?

Can linguistic syntax improve PBSMT?

(Koehn et al 2003) tried to impose syntactic constituents on phrase extraction

Hierarchical Phrase structure (Chiang 2005)

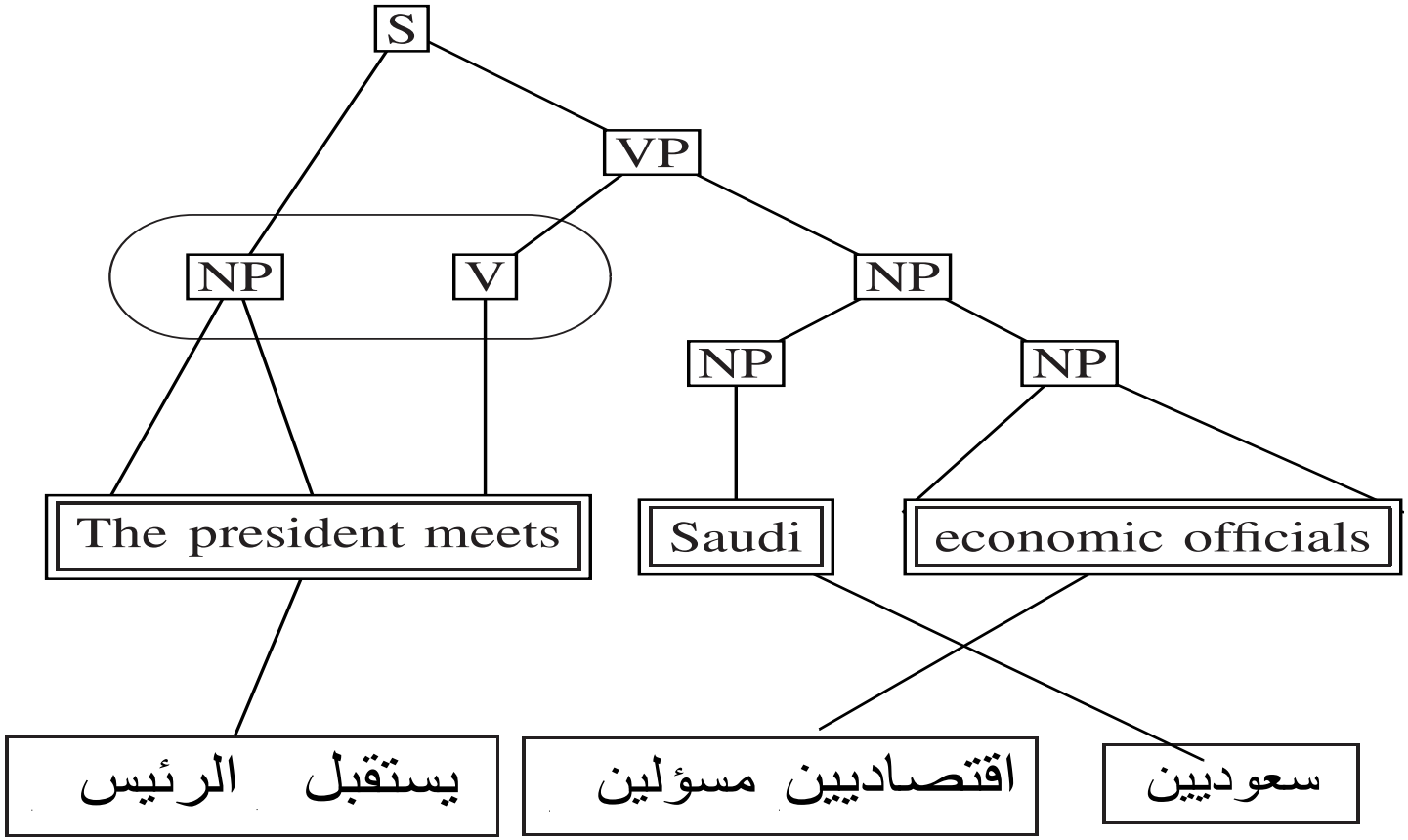
- Allows for hierarchical phrases
- Handles a range of reordering problems
- The syntax induced is not linguistically motivated.

Syntactified target phrases (Marcu et. al. 2006)

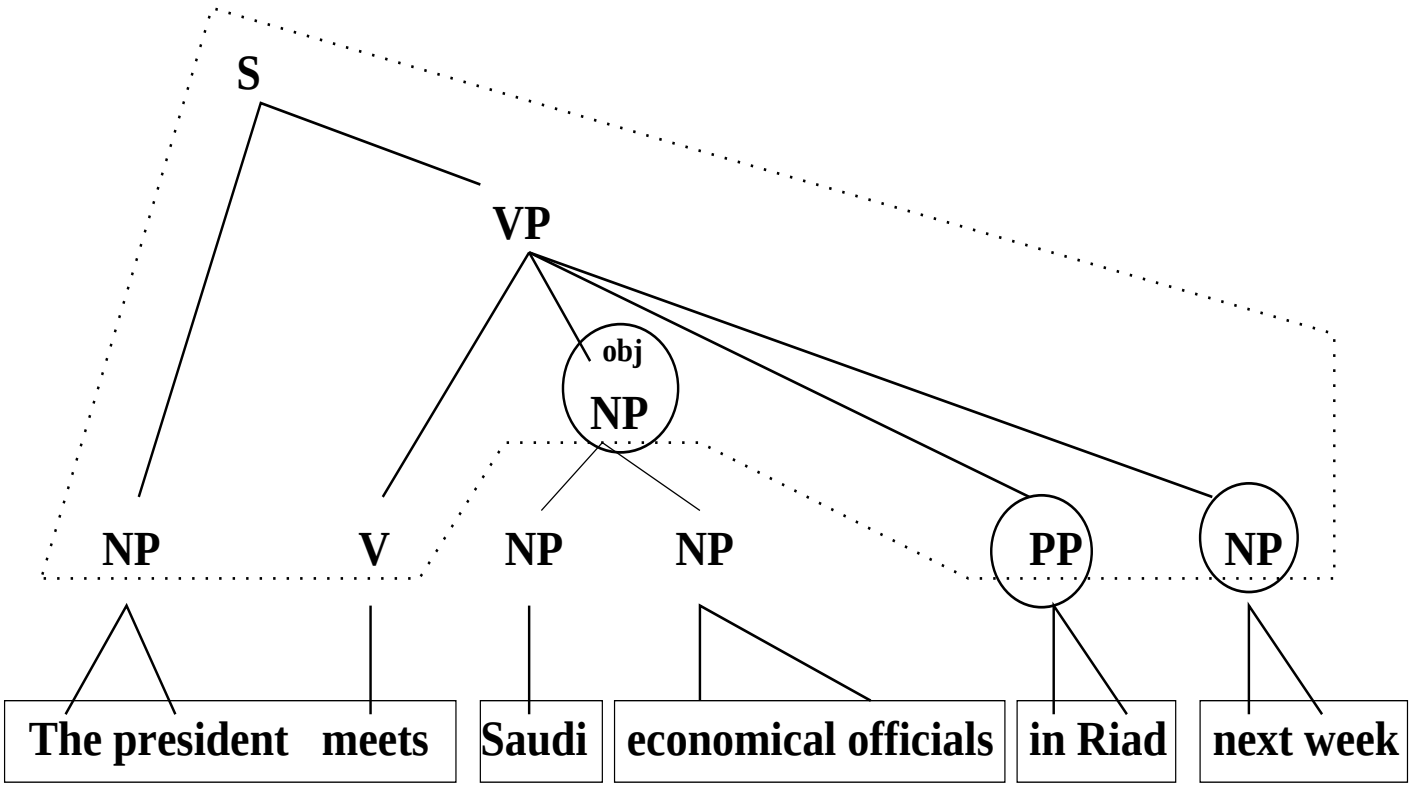
- Induces millions of xRs rules from parallel corpus
- Mismatch between constituent (xRs) and phrase
- Subtrees for phrases: leads to spurious ambiguity in phrase table

Do subtrees/constituents fit well with phrases?

Do subtrees/constituents fit well with phrases?



Spurious Ambiguity:



Do subtrees/constituents fit well with phrases?

Why subtress do not match SMT phrases?

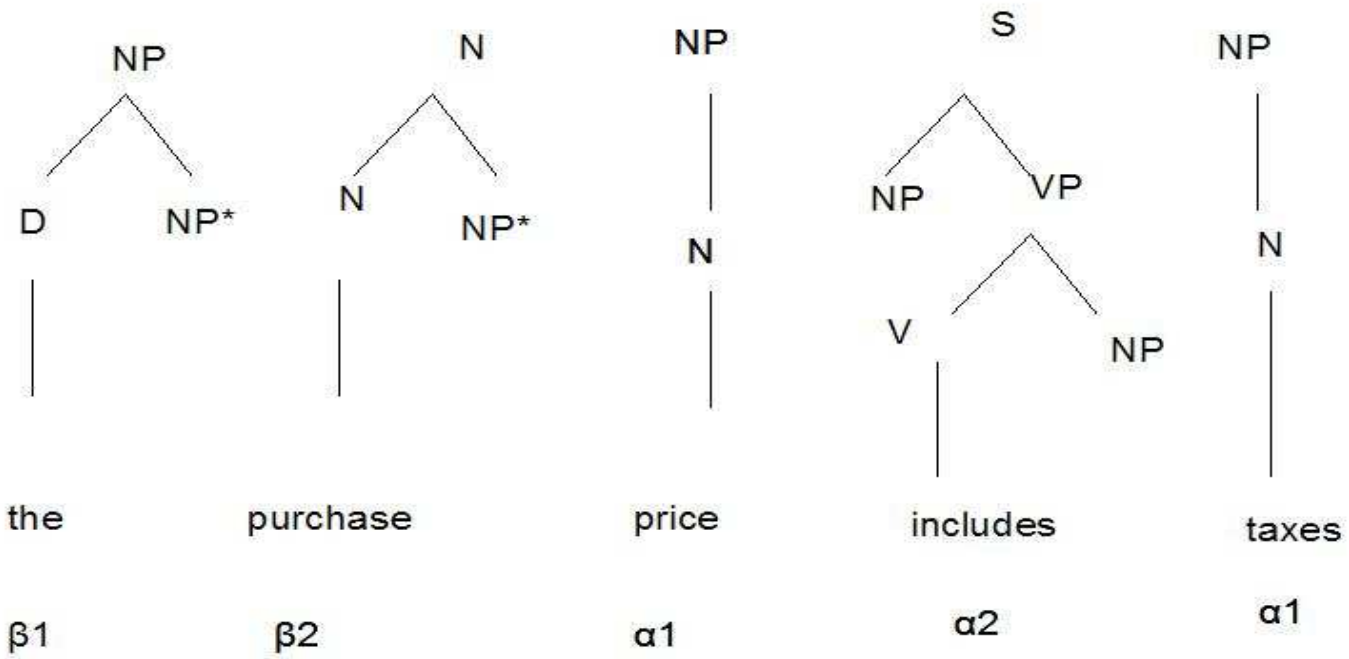
- Syntactic constituents mismatch phrase concept
- Which level of tree structure should be incorporated ?
- This leads to spurious ambiguity

Can linguistic syntax improve PBSMT?

Trees/constituents do NOT fit well with phrases

What syntax does fit then ?

Lexical Syntax (Supertags) Matches Phrases



Lexical Syntax (Supertags)

Linguistics offers lexical-syntax (Supertags):

- Lexicalized Tree Adjoining Grammar (LTAG) : (Joshi & Schabes, 1992) & (Srinivas & Joshi, 1999)
- Combinatory Categorical Grammar (CCG) (Steedman,2000)

Rich lexical categories

- Localizing syntactic dependencies
- Representing predicate argument constraints on the word level
- Markovian language model on the sequence produce almost parsing
- Handful of Combination Operators are used to construct dependency tree

Lexical Syntax (Supertags)

Linguistics offers lexical-syntax (Supertags):

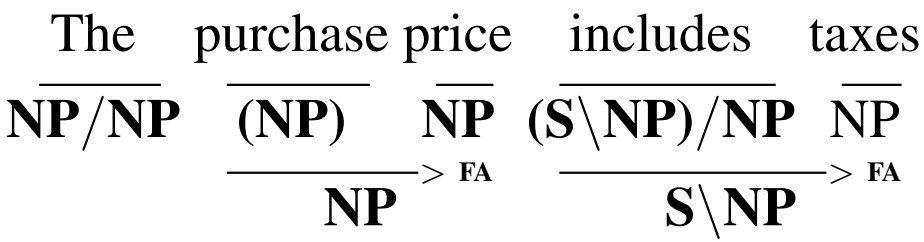
- Lexicalized Tree Adjoining Grammar (LTAG) : (Joshi & Schabes, 1992) & (Srinivas & Joshi, 1999)
- Combinatory Categorical Grammar (CCG) (Steedman,2000)

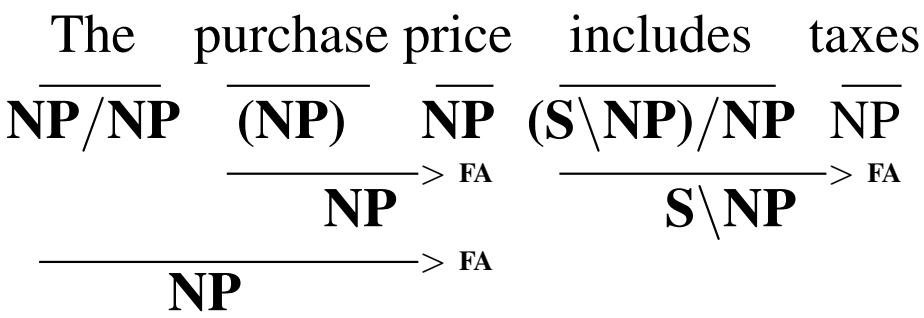
Rich lexical categories

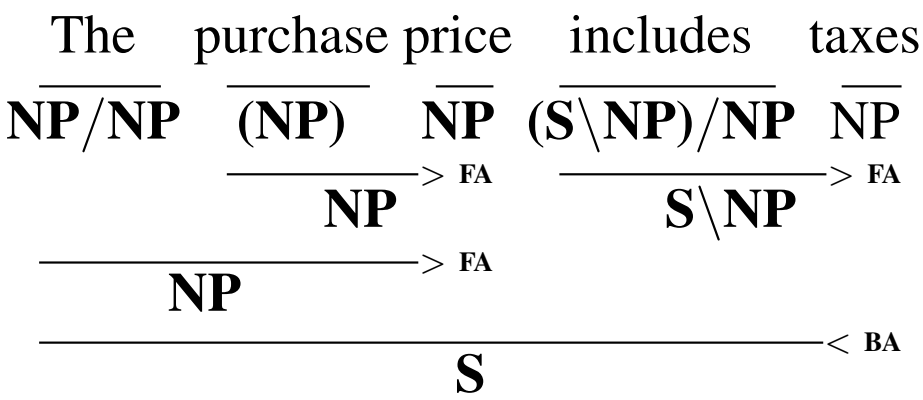
- Localizing syntactic dependencies
- Representing predicate argument constraints on the word level
- Markovian language model on the sequence produce almost parsing
- Handful of Combination Operators are used to construct dependency tree

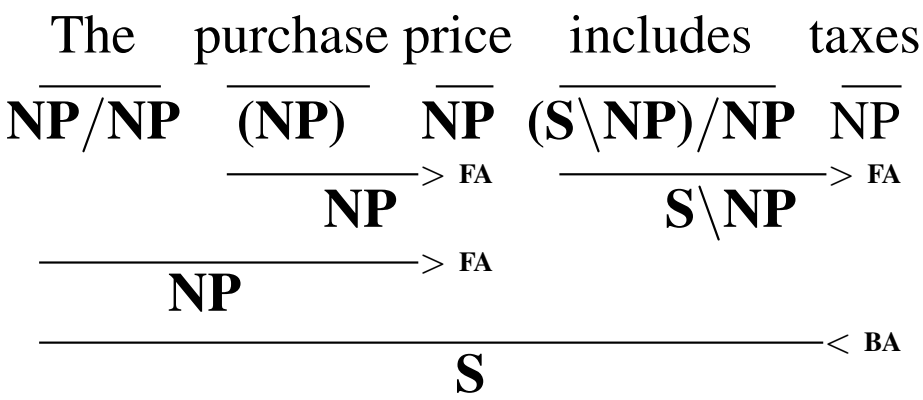
The purchase price includes taxes

The	purchase price	includes	taxes
NP/NP	**(NP)**	**(S\NP)/NP**	NP
	NP		









Lexical Syntax for SMT

Two levels of support:

- Supertagged TM & LM
- Fully incremental parsing

Outline

Syntactically Lexicalized Phrase Based SMT

Can linguistic syntax improve the output of Phrase-based SMT systems?

- Which syntax could fit with PBSMT ?
- Lexical Syntax : LTAG/CCG Supertags
- Supertags improve the performance of state-of-the-art PBSMT system on large data sets:
- Arabic-to-English NIST'05
- German-to-English shared task 07

Baseline PBSMT vs Supertags PBSMT

Baseline PBSMT

- Many candidate phrases
- Not constrained enough
- N-gram LM can not choose best candidates

Supertags PBSMT

- Many candidate phrases
- Syntactically Constrained Phrases
- Further sophisticated techniques could choose best candidates

Supertagged PBSMT Model

Supertags PBSMT: Noisy Channel Model

$$\arg\max_{t} \sum_{ST} P(s \mid t, ST) P_{ST}(t, ST) \approx$$

$$\arg\max_{t,ST} P(s \mid t, ST) P_{ST}(t, ST) \approx$$

$$\arg\max_{\sigma,t,ST} \overbrace{P(\phi_s \mid \phi_{t,ST})}^{TM\ \ w.sup.tags} \overbrace{P(O_s \mid O_t)^{\lambda_o}}^{distortion} \overbrace{P_{ST}(t, ST)}^{LM\ \ w.sup.tags}$$

Supertagged PBSMT Model

Supertags PBSMT: Log-Linear Model

$$t^* = \arg\max_{t,\sigma,ST} \prod_{f \in F} H_f(s,t,\sigma,ST)^{\lambda_f}$$

- Log-linear model representation
- Added features for supertags

Supertagged PBSMT Model

Supertags Language Model

$$P(t, ST) = \prod_{i=1}^{n} P(st_i | st_{i-4}^{i-1}) P(t_i | st_i)$$

- Log-linear Language Model for Supertags
- 5-gram Markov Language Model over supertags sequence

Supertagged PBSMT Model

Supertagged Phrase Translation Probability

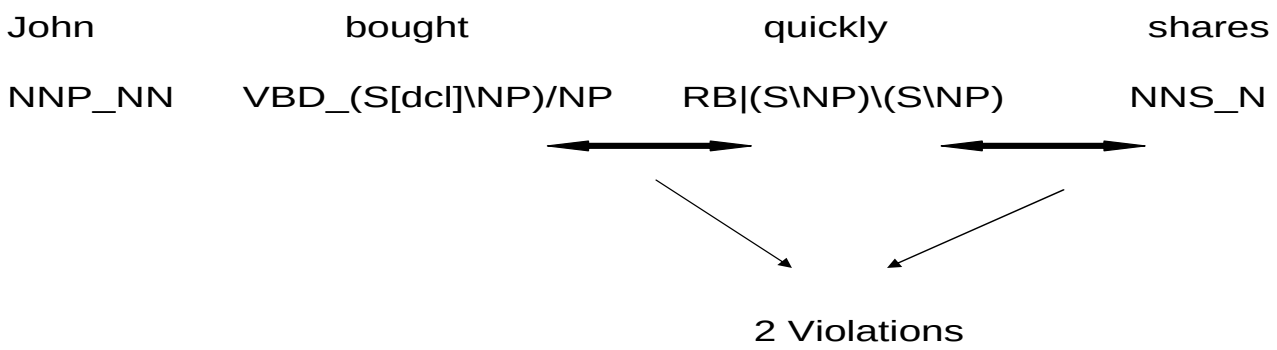
$$P(\phi_s \mid \phi_{t,ST}) \approx \prod_{\langle s_i, t_i ST_i \rangle \in (\phi_s \times \phi_{t,ST})} P(s_i \mid t_i, ST_i)$$

$$P(\phi_{t,ST} \mid \phi_s) \approx \prod_{\langle s_i, t_i ST_i \rangle \in (\phi_s \times \phi_{t,ST})} P(t_i, ST_i \mid s_i)$$

- Phrase translation probability and its reverse
- Generate target words and supertags simultaneously

LMs with Global Grammaticality Measures

- Log-linear feature
- Smoothing factor for supertags LM
- Number of operator violations



LMs with Global Grammaticality Measures

He	*believes*	*in*	*what*	*he*	*said*
NP	$\mathbf{(S_{[dcl]} \backslash NP)/S_{[dcl]}}$	*PP/NP*	*NP/(S/NP)*	*NP*	*(S\NP)/NP*

The supertag of "believes" (in boldface) demands directly to its right (for "in") an "$S_{[dcl]}$" (Forward Application); however, it finds a "(in PP/NP)" instead. This counts as a single violation $V = 1$.
Note that the supertag that fits best in the given sequence for "believes" is "(S\NP)/PP".

Experimental Setup

Two Language Pairs:

- Arabic to English NIST 05
- German to English Shared Task 07

Supertaggers:

- LTAG supertaggers (XTAG and Bangalore's Maxent tagger)
- CCG supertagger (C&C tools)

Supertags:

- N-gram Language model on supertags sequence for LTAG & CCG
- Grammatical validation for CCG operators

Scalability: Larger Training Corpora

Performance on large training data

System	BLEU Score
Base-LARGE	0.4418
LTAG-LARGE	0.4600
CCG-LARGE	**0.4609**

LM with CCG Grammaticality

Adding a grammaticality factor

System	BLEU Score
Base-LARGE	0.4418
CCG-LARGE	0.4609
CCG-LARGE-GRAM	**0.4688**

German to English Results

System	BLEU Score
Base-Line	0.2704
Supertags	0.2755
Supertags no Brevity	**0.2947**

Results Analysis

Table: How CCG improves over baseline

N=50 test sentences		
Reason	#	%
Inserting verb omitted by baseline	11	%22
Better reordering	11	%22
Better word/phrase selection	5	%10
Other reasons	23	%46

Arabic to English Examples

Reference: Annan opened an internal investigation in February but cancelled it in March in preparation for a broader, independent investigation.

Baseline: Annan was to internally in February but abolished in March as a prelude to broader and independent .

Supertags: Annan conducted an internal inquiry in February but abolished in March in preparation for broader and independent .

Arabic to English Examples

Refrence: Rabat 1-14 (AFP) - A sharp debate is raging in Morocco on the freedom of the press with regard to matters connec ted personally to King Mohamed VI following the publication of articles criticizing the Moroccan monarch's income and activit ies.
Baseline: Rabat 14-1 (afp) - was a sharp controversy in morocco on press freedom in terms of topics affecting king Mohamed VI himself after publishing articles critical of the revenues of the moroccan
Supertags: Rabat 14-1 (afp) - a sharp controversy in Morocco on press freedom in respect of topics affecting king Mohamed VI personally after the publication of articles criticizing the moroccan monarch revenues.

German to English Examples

Source: Ich habe nicht fr den Bericht Mann gestimmt, denn bei allem tatschlich notwendigen Streben nach Gleichbehandlung in Beschftigung und Beruf braucht deswegen noch nicht im bereifer soweit gegangen zu werden, dassder Schutz der Freiheiten und die Achtung des Rechtsstaates dabeivllig in Vergessenheit geraten.

Reference: I have not voted for the Mann report because, while it is indeed necessary to seek equal treatment for people in employment and occupation, it is also necessary to refrain from pushing zeal to the point of abandoning all protection of freedoms and all respect for the rule of law.

Baseline: I have voted in favour of the report because , in particular , man is actually needed quest for equal treatment in employment and occupation is therefore not yet in excess of zeal went so far as to say , the protection of freedoms and respect for the rule of law is completely forgotten .

Supertags: I have not voted for the Mann report because , in fact , with all the necessary search for equal treatment in employment and occupation is therefore not yet gone so far in excess of zeal , that the protection of freedoms and respect for the rule of law is being completely forgotten .

Outline

From Supertagged to Dependency-based Language Models

Almost prsing for MT:

- A *n*-gram language model over the sequence of supertags (*'almost parsing'*).
- *'almost parsing'* for monolingual parsing
- *'almost parsing'* for bilingual parsing

What is the parsing mechanism we need for SMT?

- Suppoet long-range dependencies
- Distinguish between different translation candidates based on their role in constructing the parse structure
- Satisfy the syntactic dependencies
- Work in an incremental manner similar to SMT decoders
- Be computationally efficient to be integrated into SMT

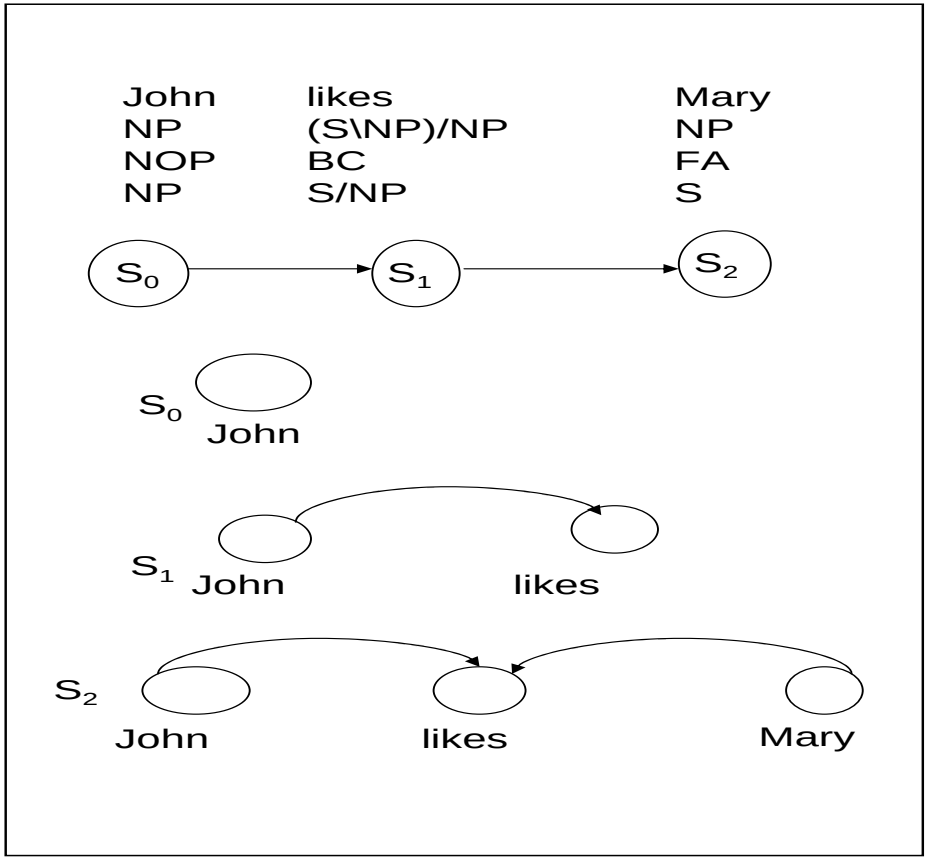
What is the parsing mechanism we need for SMT?

Our proposed IDLM differs from the related work in four major respects:

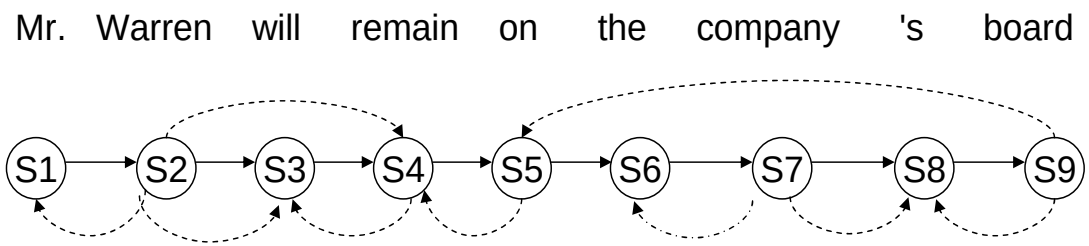
- It is based on incremental parsing that seamlessly matches the incremental nature of SMT decoders.
- It is deterministic, in the sense that it maintains a limited number of parse-states that represent possible parsing decisions at each word position. This characteristic is very important for incorporating IDLM into large-scale MT systems due to its computational efficiency.
- The grammatical representation is based on CCG structures which enable the handling of non-constituent constructions.
- The parser seeks out intermediate connected structures, unlike previous approaches which deployed dependency relations or head words to enable syntax-based probabilities into the language model.

Outline

Incremental Parsing Representation

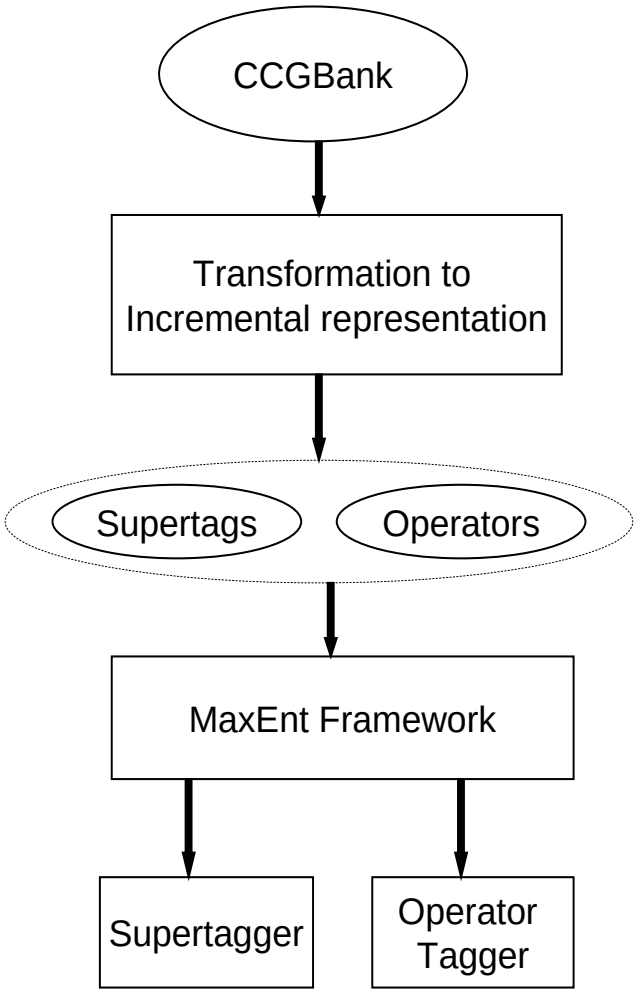


Incremental CCG

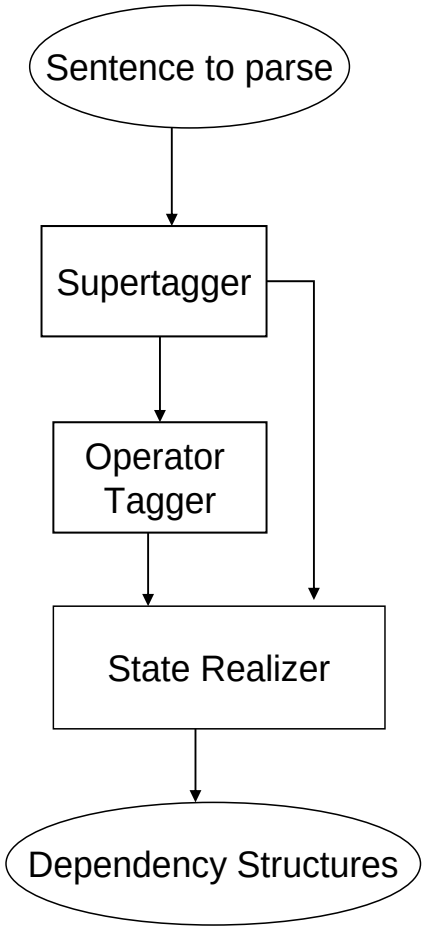


Supertag	NP/NP	NP	(S\NP)/(S\NP)	(S\NP)/PP	PP/NP	NP/NP	NP	(NP/NP)\NP	NP
Operator	NOP	FA	TRFC	FC	FC	TRFC	FA	FC	FA
State Cat.	NP/NP	NP	S/(S\NP)	(S/PP)	(S/NP)	(S/(NP\NP))/NP	(S/(NP\NP))	(S/NP)	S

Incremental Parsing Training



Incremental Parsing Runtime



Incremental Parsing Features: Apposition Handling

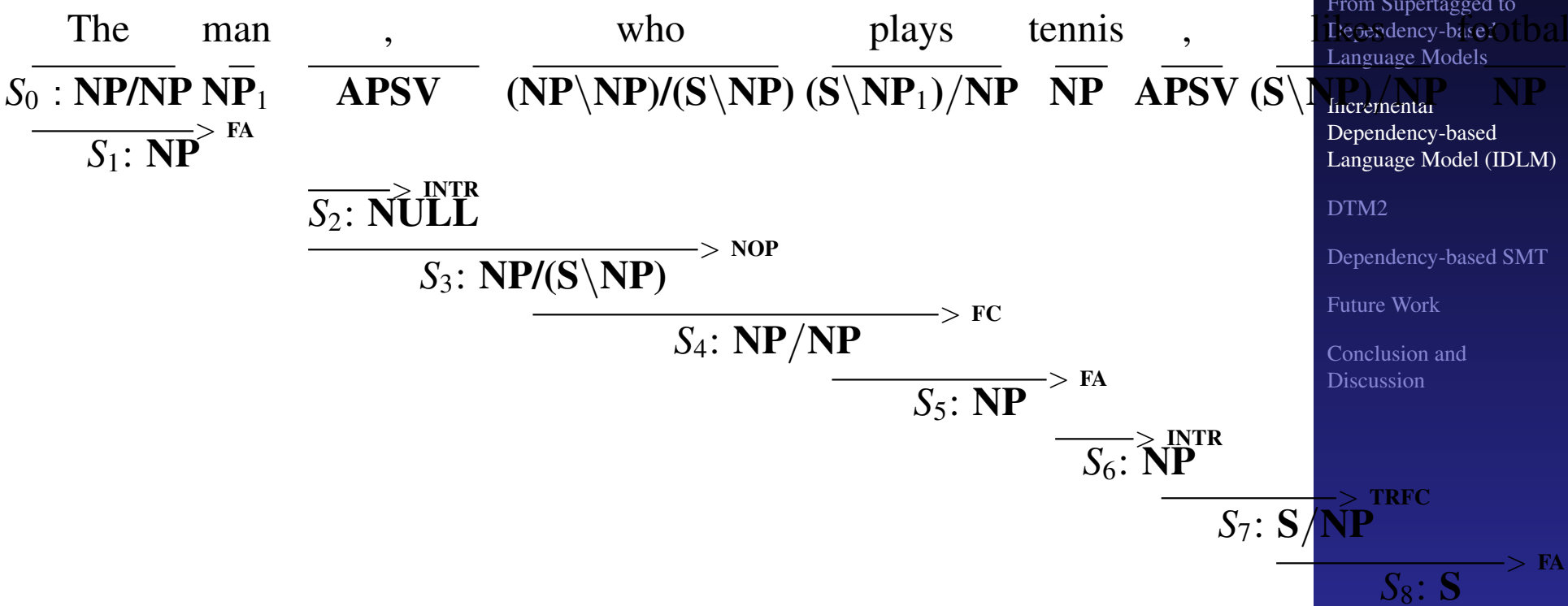


Figure: Apposition Handling.

Incremental Parsing Features: Coordination Handling

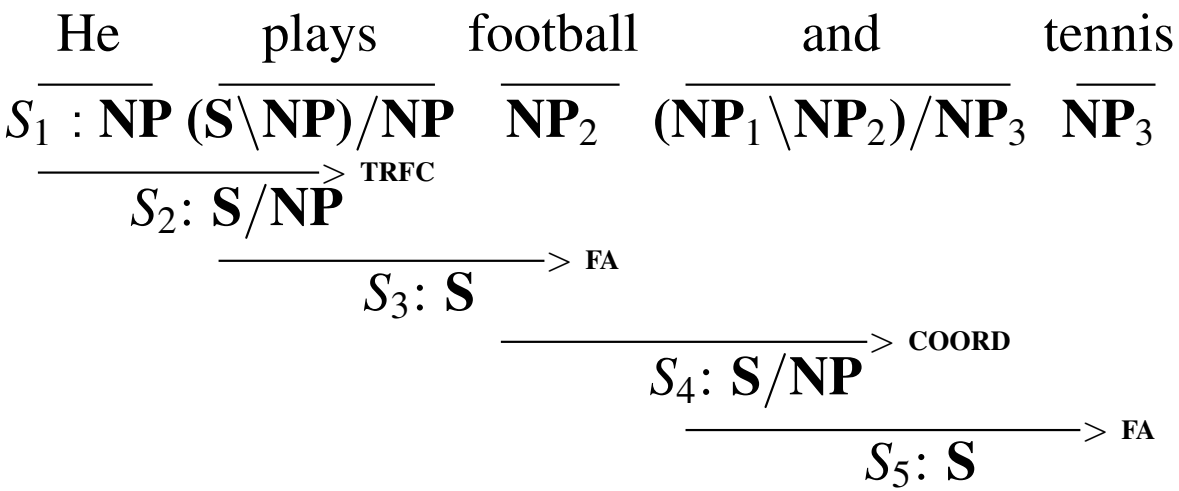


Figure: Coordination Handling.

Incremental Parsing Features: WH-movement Handling

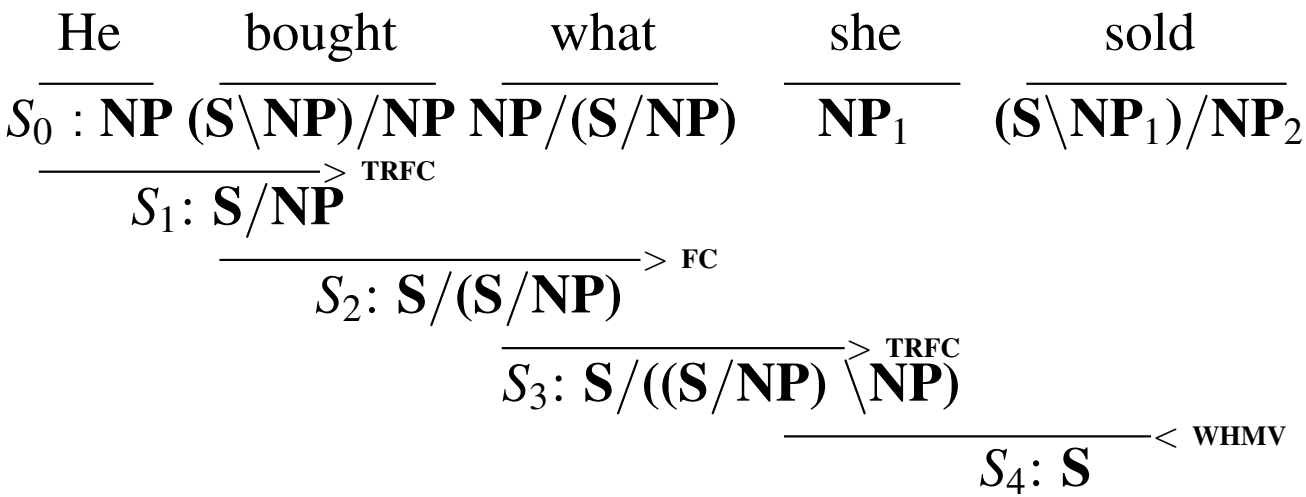


Figure: WH-movement Handling.

Incremental Parsing Evaluation

	Accuracy %	
Input	**Our tags**	**CCGbank tags**
Gold std. POS	92.39	92.46
System POS	91.7	91.81

Table: Supertagger Results.

Input/Features	**Accuracy %**
Gold standard POS and Supertags	96.73
System POS and Supertags	90.90
Preceding correct state as feature	99.22

Table: Operator Tagger Results.

Input	**F-Measure**
Gold standard POS and Supertags	87.5
System POS and Supertags	86.7

Table: Unlabeled dependency results for section 23.

Outline

MERT Estimation for log-linear Models:

- Approximiation for Maximum Entroy log-linear models
- Can handle a few number of paremeters (in order of ten)
- A bottleneck to further serious development of features-rich SMT systems
- Parameters of different components are not related

DTM2 Phrase Structues

Algnp Almrkzyp llhzb	of the X committee central of the X Party

Figure: Phrase structures in DTM2. X represents a variable in the target phrase

Outline

Dependency Direct Translation Model(DDTM)

$$P(T|S) = P_0(T,J|S)/Z \quad exp \sum_i \lambda_i \phi_i(T,J,S) \tag{1}$$

- P_0 is the prior distribution for the phrase probability
- J is the skip reordering factor for this phrase pair

DDTM Features

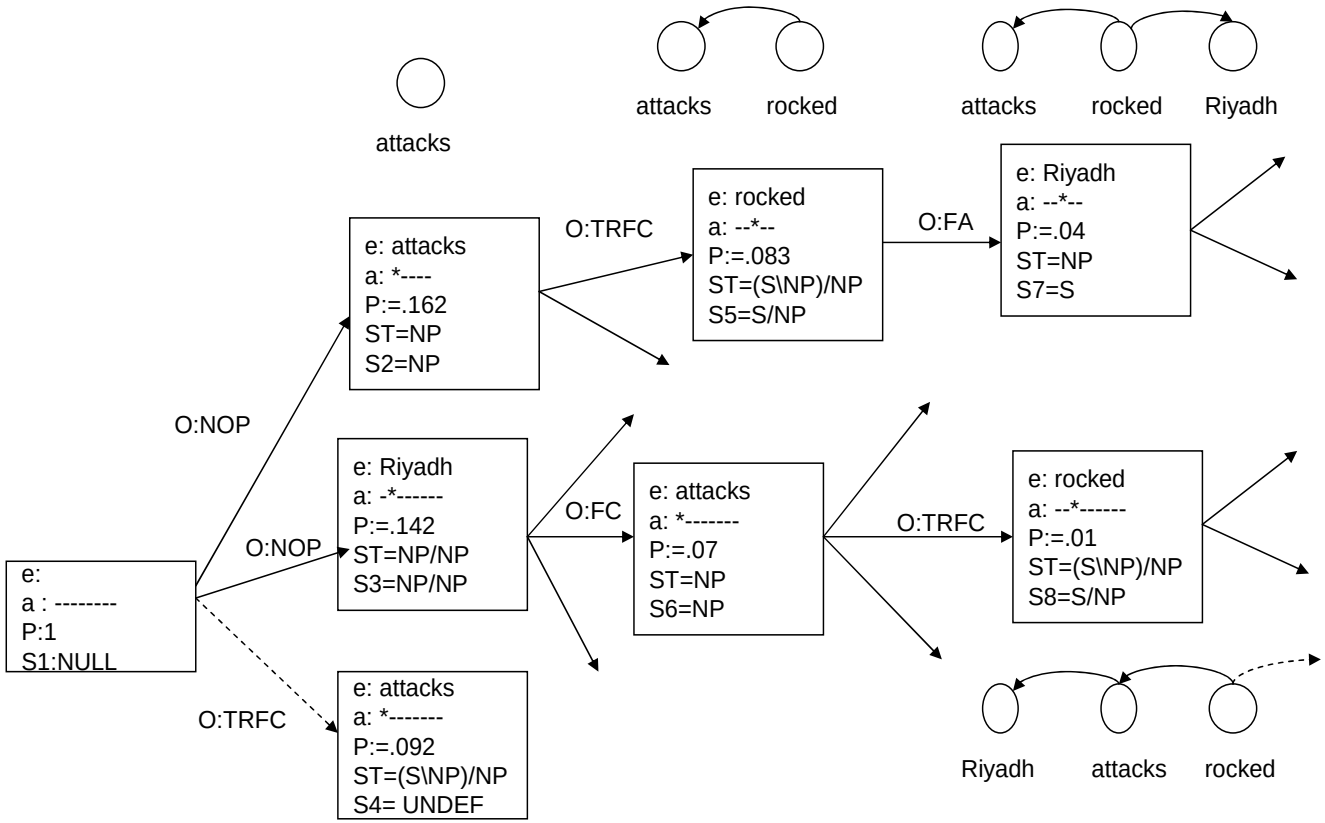
In our DDTM, we have implemented many features along with the baseline DTM2 features:

- Supertag-Word features: these features examine the target phrase words with their associated supertags.
- Supertag sequence features: these features encode n-gram supertags (equivalent to the n-gram supertags Language Model).
- Supertag-Operator features: these features encode supertags and their associated operators.
- Supertag-State features: these features encode states and supertags co-occurrence.
- State sequence features: these features encode n-gram states features and are equivalent to an n-gram states Language Model.
- Word-State sequence features: these features encode words and states co-occurrence.

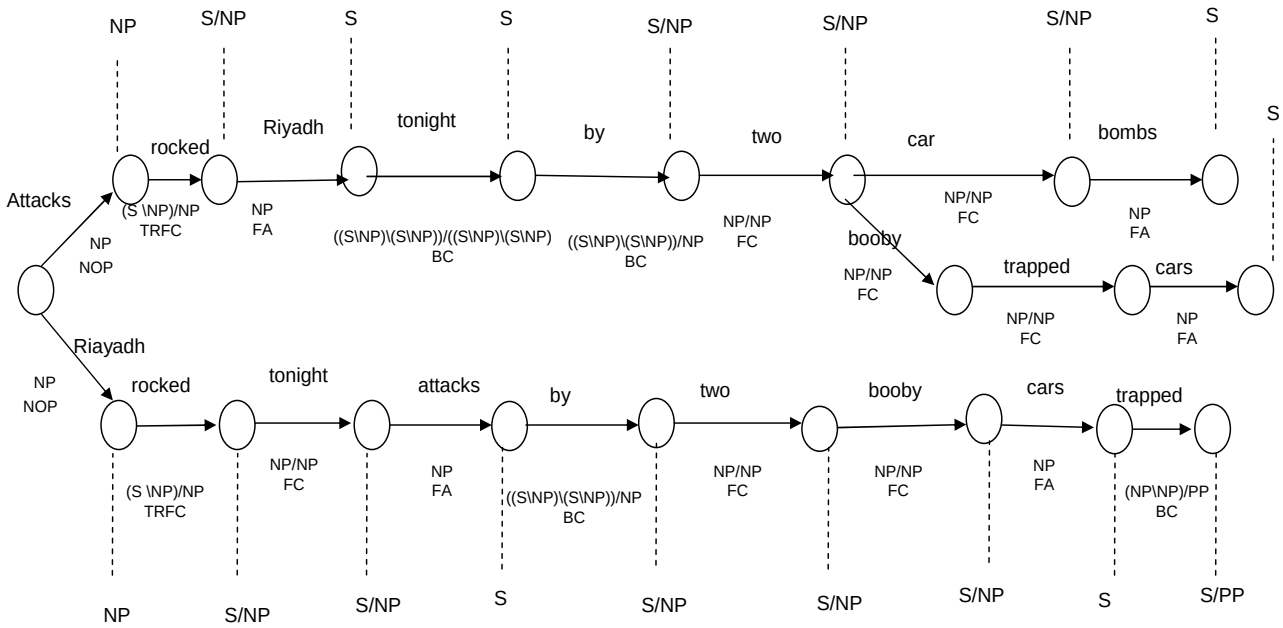
DDTM Decoder

- A beam search decoder similar to decoders used in standard phrase-based log-linear systems
- Performs incremental dependency parsing during decoding
- Supports new pruning strategies to handle the large search space

DDTM Decoder



DDTM Decoder



Experiments

- Arabic–English with 3.7M parallel sentences.
- 5-gram LM trained on English Gigaword Corpus.
- Testset: Arabic–English MT05
- Baseline is top ranked in two recent MT large scale evaluations

Evaluated Systems

- IBM-PB: IBM Phrase–based SMT baseline system.
- DTM2: the baseline Direct Translation model system.
- D-SW: examines Supertag-Word features.
- D-SLM: examines Supertag-Word features and supertags n-gram features.
- D-SO: examines Supertag-Operator features.
- D-OLM: examines operator n-gram features.
- D-SS : examines supertags and states features with parse-state construction.
- D-WS : examines words and states features with parse-state construction.
- D-SLM: examines n-gram states features with parse-state construction.
- DDTM: fully fledged system with all features that proved useful above.

Results

System	BLEU Score on MT05
IBM-PB	50.16
DTM2-Baseline	52.24
D-SW	52.28
D-SLM	52.29
D-SO	52.01
D-OLM	51.87
D-SS	52.39
D-WS	52.03
D-SLM	52.53
DDTM	52.61

Table: DDTM Results with various features.

Examples

Source: وخضع بعد ذلك لفحوصات اجراهَا احد اطباء الشرطة

Reference: *He then underwent medical examinations by a police doctor .*
Baseline: *He was subjected after that tests conducted by doctors of the police .*
DDTM: *Then he underwent tests conducted by doctors of the police .*

Example

Source: وقد هز الرياض مساء اليوم هجومان بسيارتين مفخختين

Reference: *Riyadh was rocked tonight by two car bomb attacks..*
Baseline: *Riyadh rocked today night attacks by two booby - trapped cars.*
DDTM: *Attacks rocked Riyadh today evening in two car bombs.*

Figure: DDTM provides better syntatctic structure with more concise translation.

Outline

Future Work

- Source Dependency information
- Enhance the Dependency parser accuracy
- Possible implementation of the framework into Moses.
- Extend the approach for logical semantics as well

Outline

- We introduced a novel model of supertagged Phrase-based SMT which integrates supertags into the target language model and the target side of the translation

- We introduced a novel dependency-based LM which is deterministic in that it maintains a limited number of parsing decisions at each state which. Furthermore, it is incremental in Markovian fashion similar to Phrase-based SMT decoders and it can naturally handle non-constituent constructions, being based on CCG.

- We introduced an extension to direct translation models that integrates incremental dependency parsing while retaining the linear decoding assumed in conventional Phrase–based SMT systems.

Thanks

Thanks for Listening