If I Only Had a Parser: Poor Man's Syntax for Hierarchical Machine Translation

David Vilar, Daniel Stein, Stephan Peitz and Hermann Ney
{vilar,stein,peitz,ney}@i6.informatik.rwth-aachen.de

2nd December 2010

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

1 Introduction

- Hierarchical phrase-based model
- Formally defined as a (synchronous) CFG
  - Appealing to include syntactic information
- Goal: Comparison of different previous approaches
  - Parsematch features
  - String-to-Dependency
  - Soft syntax labels (preference grammars)
- Novel approach based on phrase clustering
  - No need for linguistic tools
  - Reduction of training time

Outline

2 Related Work

- **Hierarchical phrase-based model**
  [Chiang 05], [Chiang 07]
- **Syntax extensions**
  - **Parsematch: Features depending on syntactic categories of the phrases**
    [Vilar & Stein+ 08], [Marton & Resnik 08]
  - **String-to-dependency: Reconstruct dependency tree at translation time**
    [Shen & Xu+ 08]
  - **Soft syntax labels: Extend set of non-terminals with syntactic labels**
    [Zollmann & Venugopal 06], [Venugopal & Zollmann 09]
  - **Different groups on different corpora**

3 Hierarchical Phrases

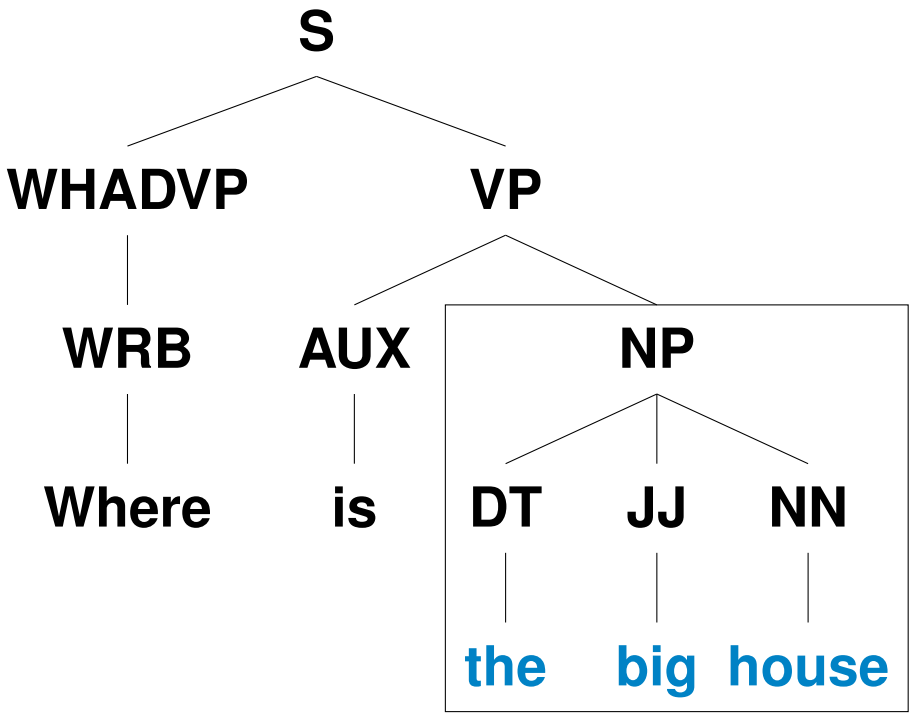
- **Formalization as a synchronous CFG**
- **Rules of the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where:**
  - **X is a non-terminal**
  - **γ and α are strings of terminals and non-terminals**
  - **$\sim$ is a one-to-one correspondence between the non-terminals of α and γ**
- **Example:**

$$X \rightarrow \langle \text{不是} \ X^{\sim 1} \ \text{的} \ X^{\sim 2}, \text{is not the } X^{\sim 2} \text{ of the } X^{\sim 1} \rangle .$$

- **Two-pass extraction process**
  - **Extract lexical phrases (as in PBT)**
  - **Extract hierarchical rules**

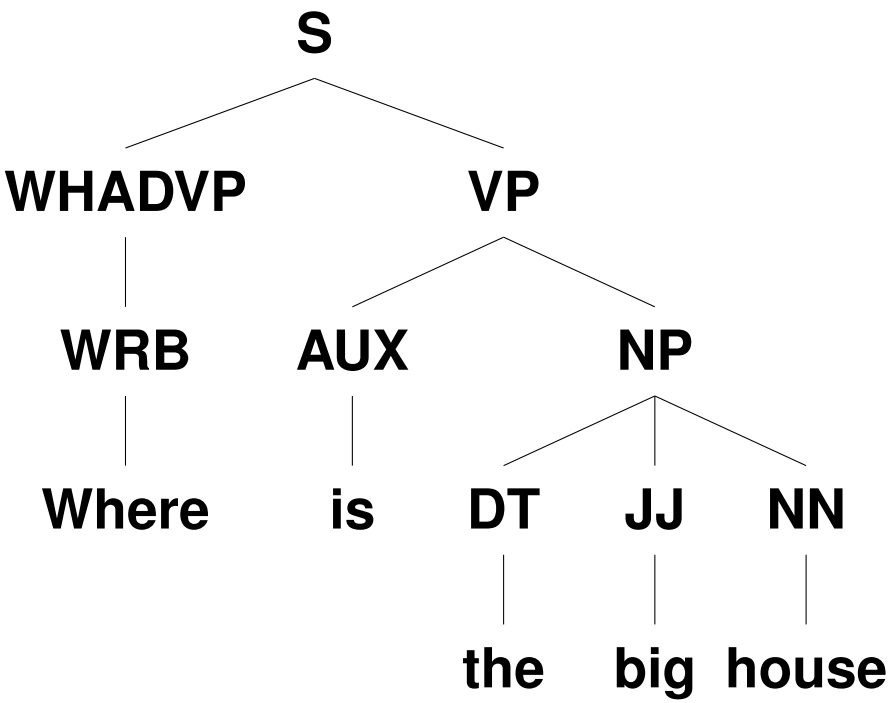
4 Syntax-based Approaches

- A phrase is valid when a node exists that completely covers all positions



Best Match Node

- Try to find the node in the parse tree that best matches the phrase
- Minimize the number of words to be deleted or added to a phrase, so that it fits the yield of a node

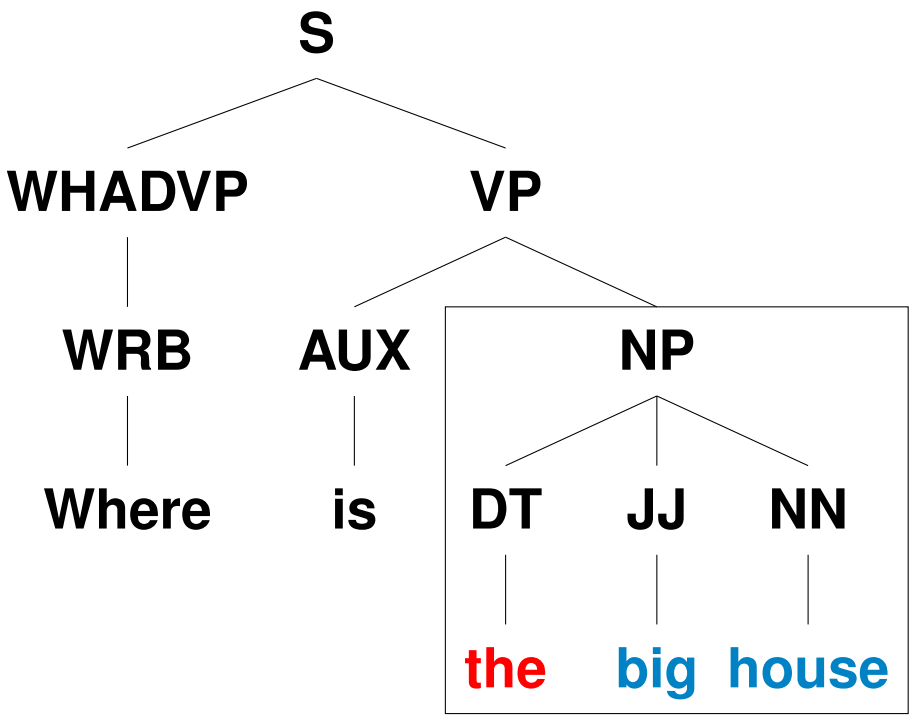


Phrases:

- big house
- is the

Best Match Node

- Try to find the node in the parse tree that best matches the phrase
- Minimize the number of words to be deleted or added to a phrase, so that it fits the yield of a node

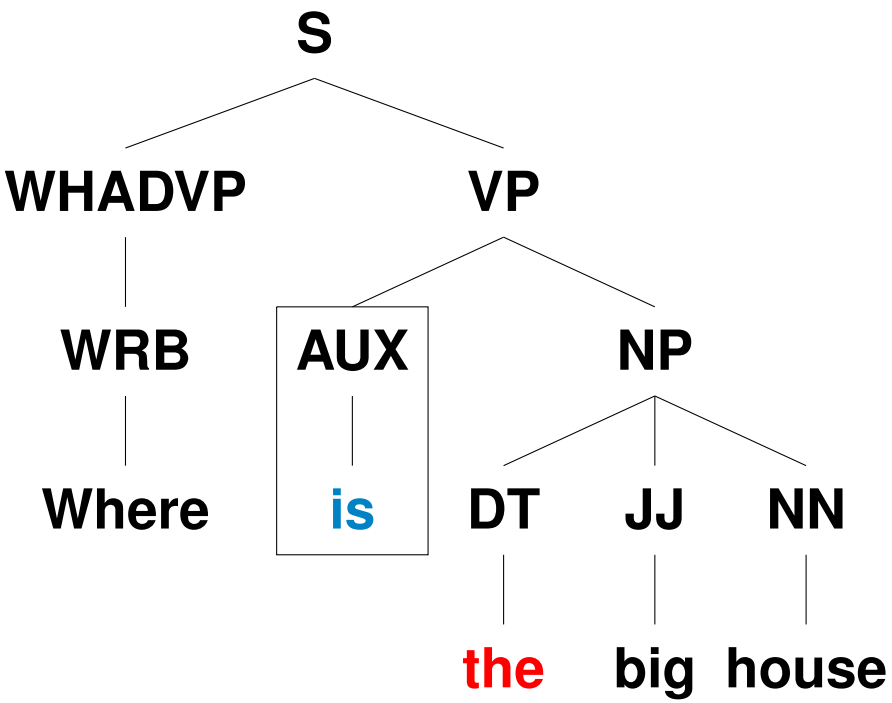


Phrases:

- big house: Node NP
- is the

Best Match Node

- Try to find the node in the parse tree that best matches the phrase
- Minimize the number of words to be deleted or added to a phrase, so that it fits the yield of a node



Phrases:

- big house: Node NP
- is the: Node AUX

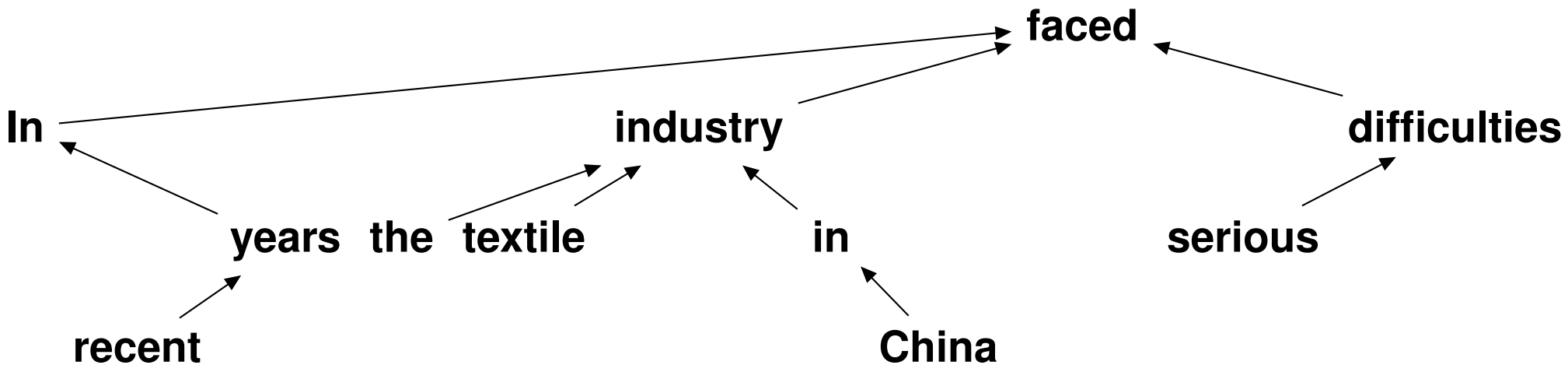
Parsematch Features

- Penalize invalid phrases
  - 0/1 feature
  - Phrase-length based features
- Simple and flexible approach
  - Applicable to source and target sides
  - Easy to adapt to other phrase-based models
- [Vilar & Stein+ 08]

String-to-Dependency

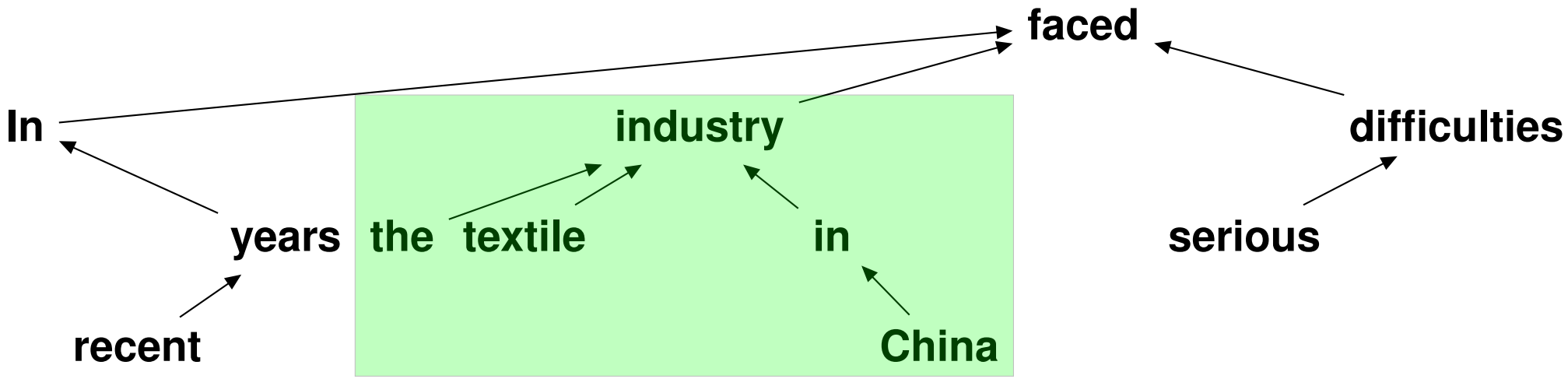
- Extract dependency structures of phrases
- (Re-)construct dependency tree while translating
- Penalize ill-formed structures
- Use additional dependency-level LMs (rescoring)
- [Shen & Xu$^+$ 08]

String-to-Dependency



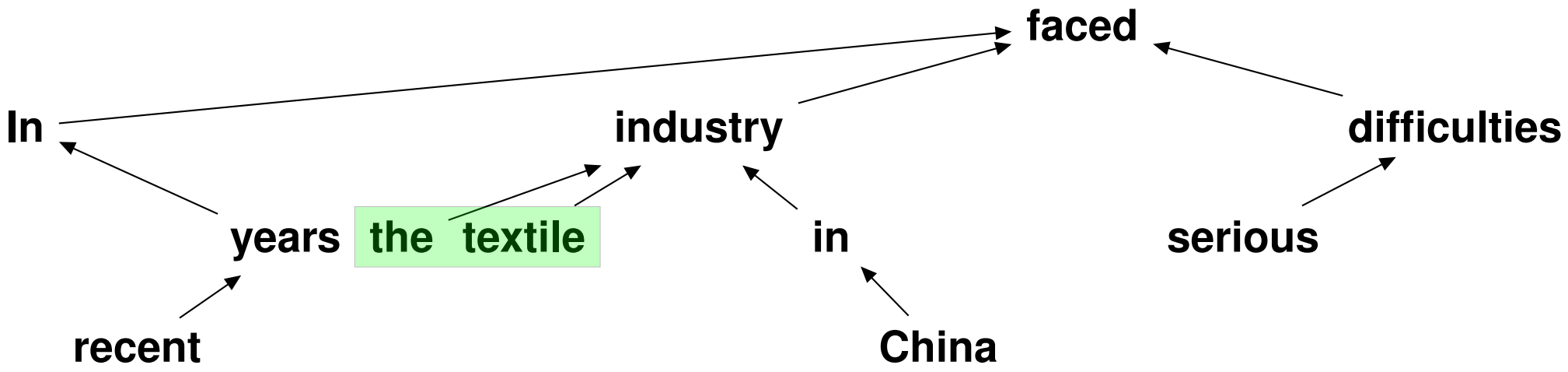
- Two types of valid dependency structures:
  - Fixed with head h
  - Floating with children C
- RWTH: do *not* restrict the rule extraction

String-to-Dependency



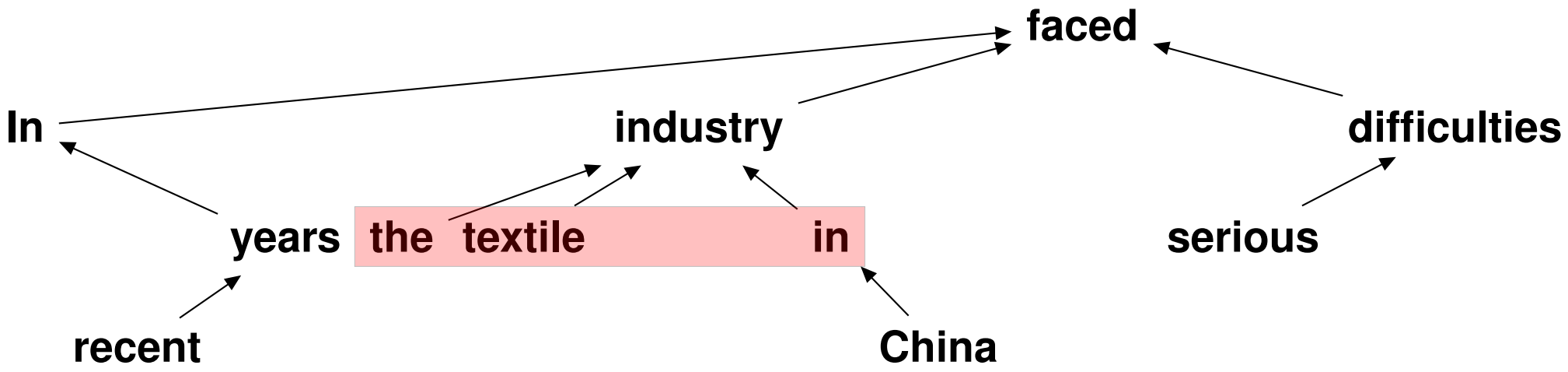
- Two types of valid dependency structures:
  - Fixed with head h
  - Floating with children C
- RWTH: do *not* restrict the rule extraction

String-to-Dependency



- Two types of valid dependency structures:
  - Fixed with head h
  - Floating with children C
- RWTH: do *not* restrict the rule extraction

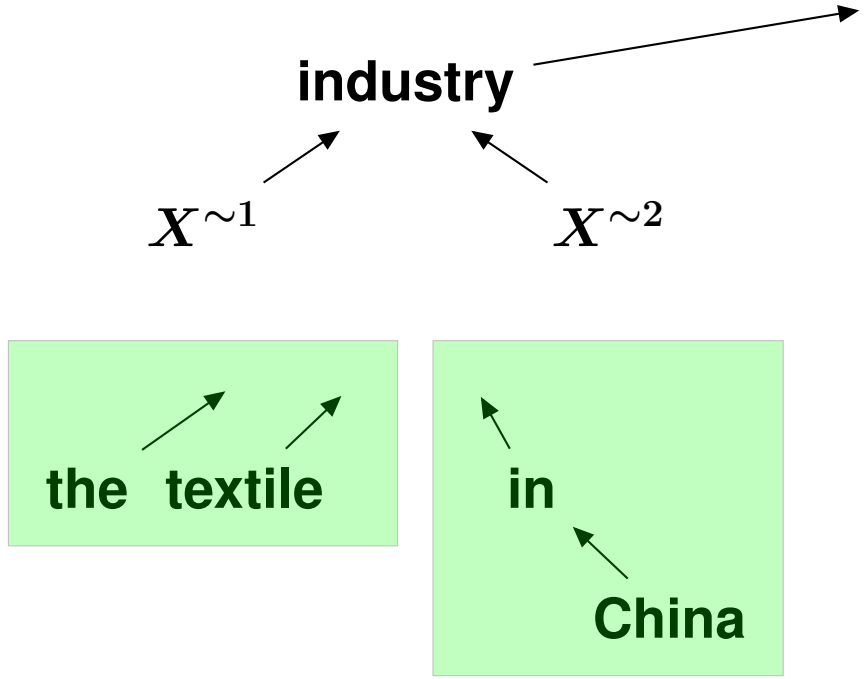
String-to-Dependency



- Two types of valid dependency structures:
  - Fixed with head h
  - Floating with children C
- RWTH: do *not* restrict the rule extraction

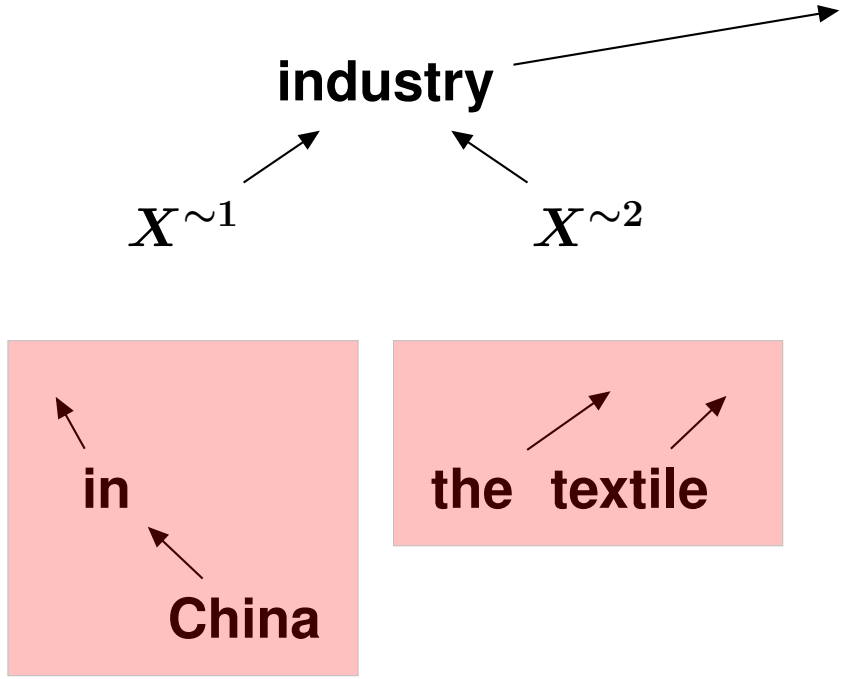
Score Computation

- Penalty for invalid dependency structures
- Penalty for merging errors
- Head probability model
- Two dependency-level LMs (right/left)



Score Computation

- Penalty for invalid dependency structures
- Penalty for merging errors
- Head probability model
- Two dependency-level LMs (right/left)



Soft Syntax Labels

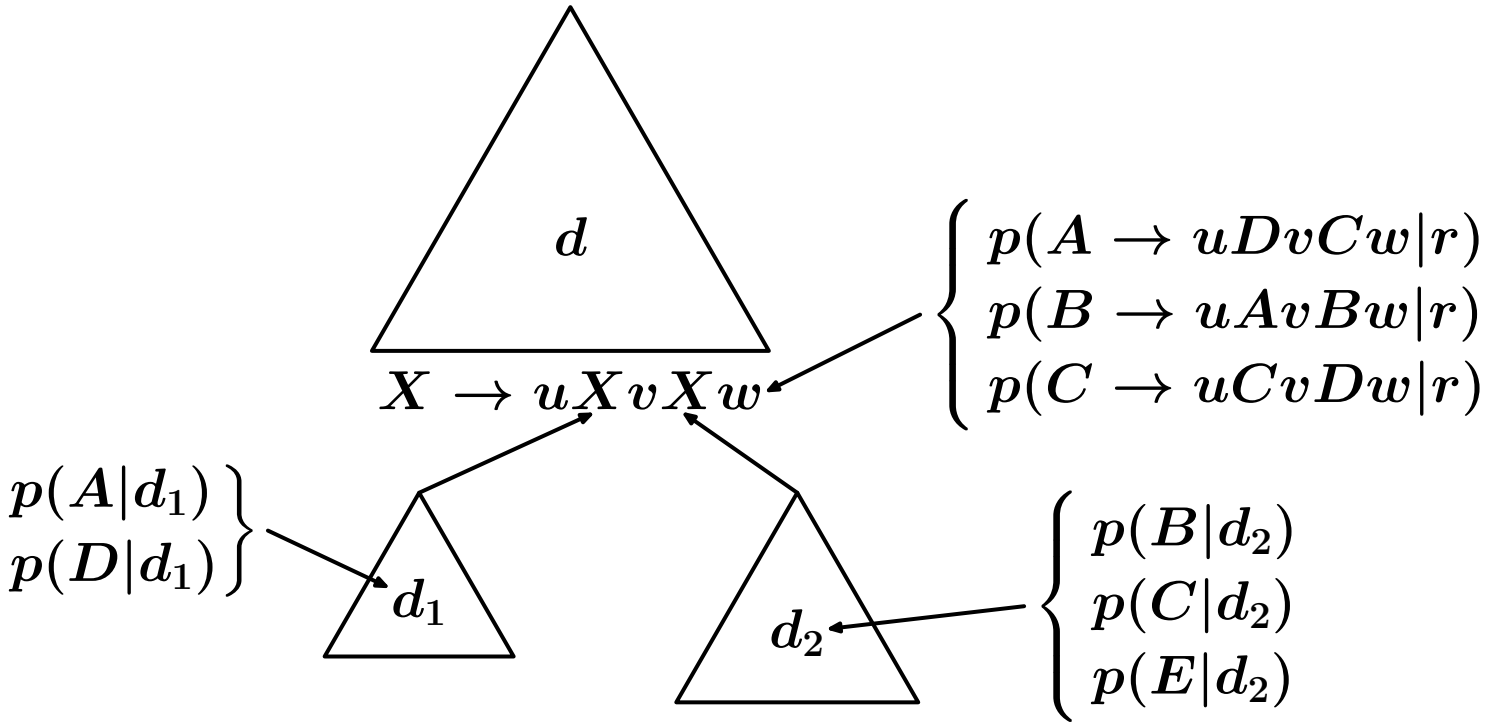
- **SAMT: Expand the set of non-terminals**

$$S \rightarrow \langle \text{Ich stimme } NP^{\sim 1} \text{ zu, I agree with } NP^{\sim 1} \rangle$$
$$SC \rightarrow \langle \text{weil andere } NP^{\sim 1} \text{ nicht } V^{\sim 2}, \text{ because others have not } V^{\sim 2}\ NP^{\sim 1} \rangle$$

- **RWTH: use labels of best match node**
- **Problem: restriction of search space**
  - **Leave rules unchanged for the parsing process**
  - **Consider the new non-terminals in a probabilistic way**
  - **Compute a "well-formedness" probability of the translation**
- **[Zollmann & Venugopal 06], [Venugopal & Zollmann 09]**

Soft Syntax Labels

Factors intervening in the computation of $p_{syntax}(d)$



5 Poor Man's Syntax

- **What is the role of the non-terminals?**

$$S \rightarrow \langle u NP^{\sim 1} v PP^{\sim 2} w, x NP^{\sim 1} y PP^{\sim 2} z \rangle$$

$NP \rightarrow \langle$ **dieser Forderung, this request** $\rangle$
$NP \rightarrow \langle$ **der schnelle braune Fuchs, the quick brown fox** $\rangle$
$NP \rightarrow \langle$ **der faule Hund, the lazy dog** $\rangle$
$\ldots$

$PP \rightarrow \langle$ **mit** $NP^{\sim 1}$**, with** $NP^{\sim 1} \rangle$
$PP \rightarrow \langle$ **über** $NP^{\sim 1}$**, over** $NP^{\sim 1} \rangle$
$\ldots$

5 Poor Man's Syntax

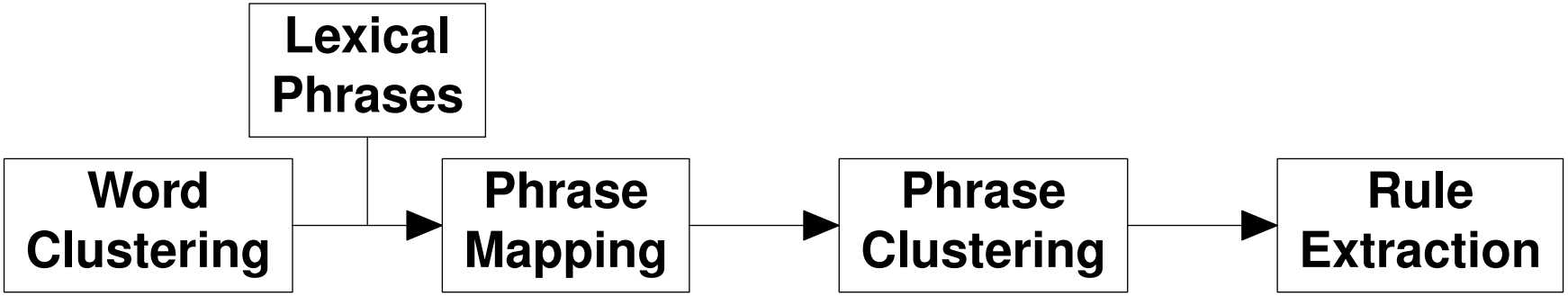
- **What is the role of the non-terminals?**

$$C2 \rightarrow \langle u \mathrm{C0}^{\sim 1} v \mathrm{C1}^{\sim 2} w, x \mathrm{C0}^{\sim 1} y \mathrm{C1}^{\sim 2} z \rangle$$

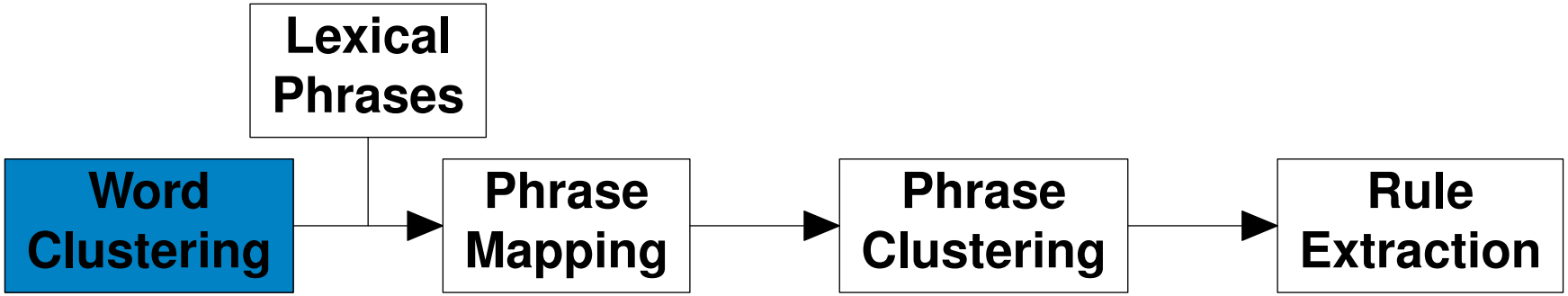
$\mathrm{C0} \rightarrow \langle$ **dieser Forderung, this request** $\rangle$
$\mathrm{C0} \rightarrow \langle$ **der schnelle braune Fuchs, the quick brown fox** $\rangle$
$\mathrm{C0} \rightarrow \langle$ **der faule Hund, the lazy dog** $\rangle$
$\ldots$

$\mathrm{C1} \rightarrow \langle$ **mit** $\mathrm{C0}^{\sim 1}$**, with** $\mathrm{C0}^{\sim 1} \rangle$
$\mathrm{C1} \rightarrow \langle$ **über** $\mathrm{C0}^{\sim 1}$**, over** $\mathrm{C0}^{\sim 1} \rangle$
$\ldots$

Poor Man's Syntax



Poor Man's Syntax



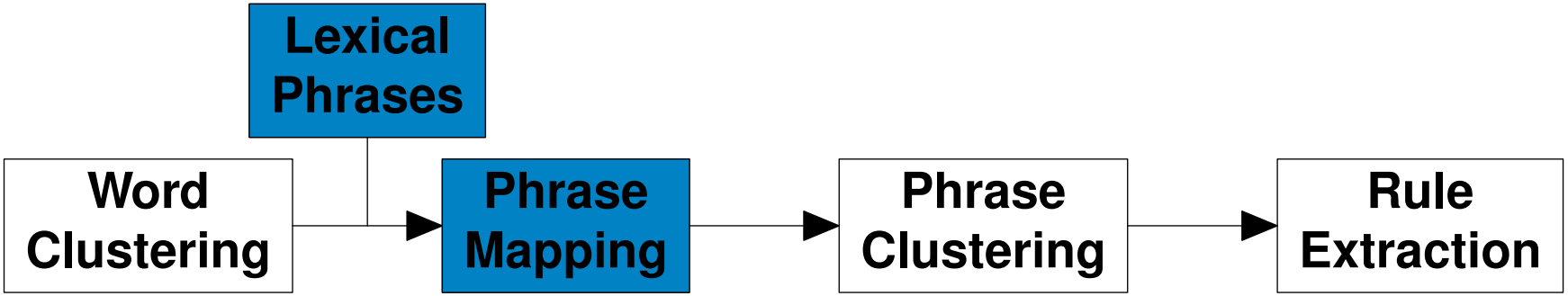
Source

...	
不是	Source class 5
美国	Source class 5
的	Source class 3
核心	Source class 3
国家利益	Source class 2
...	

Target

...	
is	Target class 3
not	Target class 5
the	Target class 4
core	Target class 2
national	Target class 4
interests	Target class 2
of	Target class 3
united	Target class 4
states	Target class 2
...	

Poor Man's Syntax



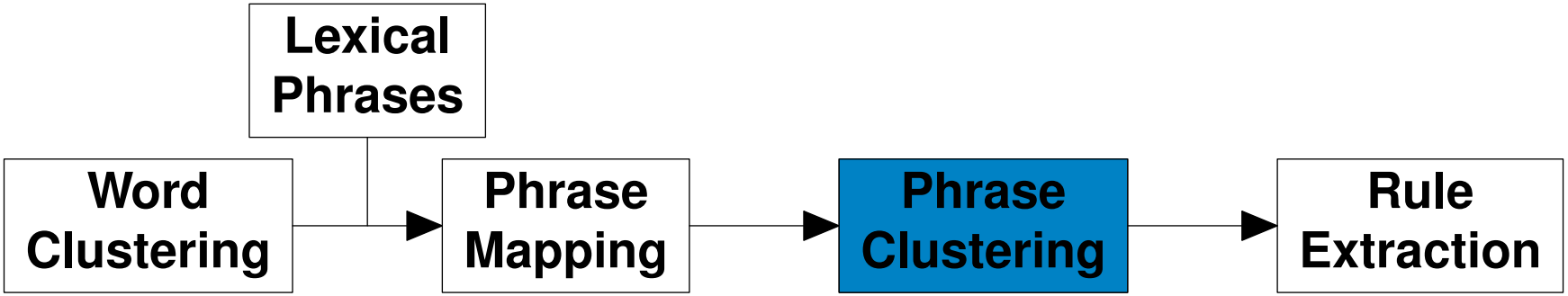
Lexical phrases

```
...
美国  # united states
...
不是 美国 的 核心 国家利益  #
      is not the core national
      interest of the united
      states
...
核心  国家利益  # core
      national interest
...
```

Mapped phrases

```
...
SC5 # TC4 TC2
...
SC5 SC5 SC3 SC3 SC2 #
      TC3 TC5 TC4 TC2
      TC4 TC2 TC3 TC4
      TC4 TC2
...
SC3 SC2 # TC2 TC4 TC2
...
```

Poor Man's Syntax



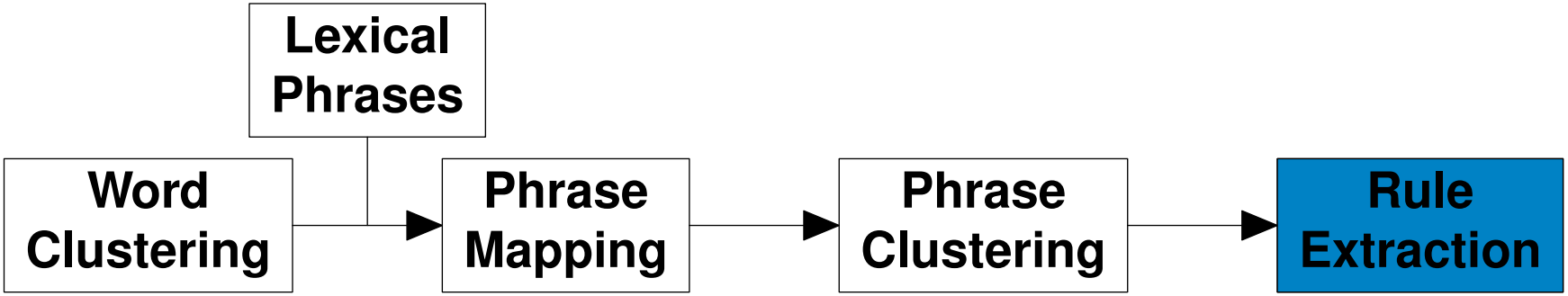
Features

```
...
SC5 # TC4 TC2
...
SC5 SC5 SC3 SC3 SC2 #
      TC3 TC5 TC4 TC2
      TC4 TC2 TC3 TC4
      TC4 TC2
...
SC3 SC2 # TC2 TC4 TC2
...
```

Clustered phrases

```
...
美国  # united states                    Class 14
...
不是 美国 的 核心 国家利益  #  Class 8
      is not the core national
      interest of the united
      states
...
核心  国家利益  # core                   Class 16
      national interest
...
```

Poor Man's Syntax



$X \rightarrow \langle$ 不是 美国 的 核心 国家利益 , **is not the core national interests of the united states** $\rangle$

$\Downarrow$ $\Downarrow$ $\Downarrow$ $\Downarrow$ $\Downarrow$

$\mathrm{C8} \rightarrow \langle$ 不是 $\mathrm{C14}^{\sim 1}$ 的 $\mathrm{C16}^{\sim 2}$, **is not the** $\mathrm{C16}^{\sim 2}$ **of the** $\mathrm{C14}^{\sim 1}$ $\rangle$

Clustering Details

- Word clustering using `mkcls` [Och 99]
  - ML on a bigram decomposition
- Phrase clustering using Cluto [Zhao & Karypis 03]
  - Standard settings, maximize cosine distance between clusters
- Cluster sizes: 5 word classes, 20 phrase classes
  - Costly experiments: clustering, phrase extraction, MERT
  - Non-exhaustive search for best number of classes
- Much more efficient that parsing the trainig data:
  - Clustering: 20 hours
  - Parsing: 2000 sentences per hour

6 Experimental Results

- NIST'08 CE Task

		Chinese	English
Train	Sentences	3 030 696	
	No. of Words	77 456 152	81 002 954
	Vocabulary	83 128	213 076
	Singletons	21 059	95 544
Dev (NIST '06)	Sentences	1 664	
	No. of Words	42 930	172 324
	Vocabulary	6 387	17 202
	OOVs	1 871	50 353
Test (NIST '08)	Sentences	1 357	
	No. of Words	36 114	149 057
	Vocabulary	6 418	17 877
	OOVs	1 375	43 724

Experimental Results

- NIST'08 CE Task

	BLEU	TER
Hierarchical baseline	24.0±0.9	68.4±0.6
Parsematch	24.4	67.9
String-to-dependency	24.6	66.7
Soft syntax labels	25.0	67.2
Parsematch + dependency	24.6	67.6
Syntax labels + parsematch	25.3	67.3
Syntax labels + dependency	25.4	66.7
Syntax labels + parsematch + dependency	25.1	66.4
Poor man's syntax	24.8	66.9

- Consistent results for NIST'06 (dev)
- Additional results in this IWSLT evaluation

Experimental Results

- **NIST'08 CE Task**

	BLEU	TER
Hierarchical baseline	24.0±0.9	68.4±0.6
Parsematch	24.4	67.9
String-to-dependency	24.6	66.7
Soft syntax labels	25.0	67.2
Parsematch + dependency	24.6	67.6
Syntax labels + parsematch	25.3	67.3
Syntax labels + dependency	25.4	66.7
Syntax labels + parsematch + dependency	25.1	66.4
Poor man's syntax	24.8	66.9

- **Consistent results for NIST'06 (dev)**
- **Additional results in this IWSLT evaluation**

7 Conclusions

- Compared different syntax enhancements
  - Same conditions (data and implementation)
  - Different effects on BLEU and TER
  - Combination helps
- New method: Poor Man's Syntax
  - Captures the structural advantages of soft syntax labels
  - No need for linguistic tools
  - Useful for under-resourced languages
- Future work
  - Further clustering methods
  - Context dependence
- Open source implementation (Jane v1.1)

If I Only Had a Parser: Poor Man's Syntax for Hierarchical Machine Translation

David Vilar, Daniel Stein, Stephan Peitz and Hermann Ney
{vilar,stein,peitz,ney}@i6.informatik.rwth-aachen.de

2nd December 2010

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

References

[Chiang 05] D. Chiang: A Hierarchical Phrase-Based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 263–270, Ann Arbor, Michigan, USA, June 2005. 4

[Chiang 07] D. Chiang: Hierarchical phrase-based translation. *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, June 2007. 4

[Marton & Resnik 08] Y. Marton, P. Resnik: Soft Syntactic Constraints for Hierarchical Phrased-Based Translation. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1003–1011, Columbus, Ohio, June 2008. 4

[Och 99] F.J. Och: An efficient method for determining bilingual word classes. *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 8–12, June 1999. 16

[Shen & Xu$^+$ 08] L. Shen, J. Xu, R. Weischedel: A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. *Proceedings of the 46th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, pp. 577–585, Columbus, Ohio, June 2008. 4, 9

[Venugopal & Zollmann 09] A. Venugopal, A. Zollmann: Preference Grammars: Softening Syntactic Constraints to Improve Statistical Machine Translation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 236–244, Boulder, Colorado, USA, June 2009. 4, 12

[Vilar & Stein+ 08] D. Vilar, D. Stein, H. Ney: Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation. *International Workshop on Spoken Language Translation*, pp. 190–197, Waikiki, Hawaii, Oct. 2008. 4, 8

[Zhao & Karypis 03] Y. Zhao, G. Karypis: Clustering in Life Sciences. *Functional Genomics: Methods and Protocols*, Vol. 224, pp. 183–218, 2003. 16

[Zollmann & Venugopal 06] A. Zollmann, A. Venugopal: Syntax Augmented Machine Translation Via Chart Parsing. *Proceedings of the Workshop on Statistical Machine Translation*, pp. 138–141, New York City, New York, USA, June 2006. 4, 12