Grammatical Machine Translation

Stefan Riezler and **John T. Maxwell III**
Palo Alto Research Center

Irena Dotcheva

Overview

- Introduction
- Extracting F-Structure Snippets
- Extracting Transfer Rules
- Parsing-Transfer-Generation
- Statistical Models and Training
- Experimental Evaluation
- Discussion
- Conclusion
- References

Introduction

- Phrase-based translation

  - lacks a mechanism to learn long-distance dependencies

  - unable to generalize to unseen phrases that share non-overt linguistic information

- Recent work deploys grammar-based statistical parsers into phrase-based SMT systems for:

  - *pre-ordering* source sentences (Xia & McCord 2004; Collins et al.2005)

  - or *re-ordering* translation model output by linguistically informed statistical ordering models (Ding & Palmer 2005; Quirk et al. 2005)

Introduction cont.

- Investigate contribution of grammar-based generation to dependency-based SMT

  - integrate the idea of multi-word translation units from phrase-based SMT into a transfer system for dependency structure snippets

  - use the same training and test data as phrase-based system of Koehn et al. 2003 for snippet extraction and training

  - statistical components modeled after phrase-based system of Koehn et al. 2003, weights trained by MER

→ the system feeds dependency-structure snippets into a grammar-based generator, and determines target language ordering by applying n-gram and distortion models after grammar-based generation

→ improving grammaticality, not reflecting ordering of reference translations

Extracting F-Structure Snippets

operates on the paired sentences of a sentence-aligned bilingual corpus

1. an improved word-alignment (intersecting alignment matrices for both translation directions)

2. source and target sentences are parsed (source and target LFG grammars); most similar f-structures in source and target are selected

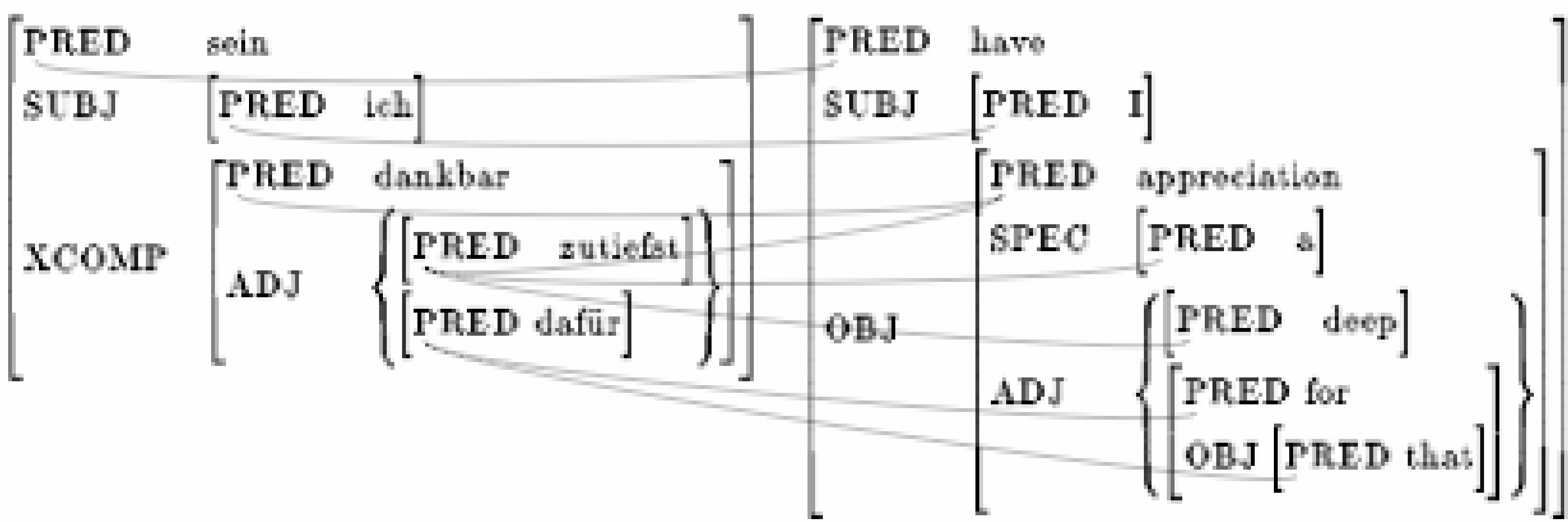
3. the many-to-many word alignment created in the first step is used to define many-to-many correspondences between the substructures of the f-structures selected in the second step

Extracting F-Structure Snippets cont.

Example sentences: *Daf¨ur bin ich zutiefst dankbar.*
I have a deep appreciation for that.

Many-to-many word alignment: *Daf¨ur* {*6 7*} *bin* {*2*} *ich* {*1*} *zutiefst* {*3 4 5*} *dankbar* {*5*}

F-structure alignment for induction of German-to-English transfer rules



Extracting Primitive Transfer Rules

First step: rule extraction directly from the alignment of f-structure units

1. maps lexical predicates

PRED(%X1, ich) ==> PRED(%X1, I)

2. maps lexical predicates and interprets subj-to-subj link as indication to map subj of source with this predicate into subject of target and xcomp of source into object of target

```
PRED(%X1,sein)              PRED(%X1,have)
 SUBJ(%X1,%X2)    ==>       SUBJ(%X1,%X2)
XCOMP(%X1,%X3)               OBJ(%X1,%X3)
```

3. maps single source f-structure into target f-structure of preposition+object units

```
                          PRED(%X1,for)
PRED(%X1, daf˙ur) ==>     OBJ(%X1,%X2)
                          PRED(%X2,that)
```

Extracting Complex Transfer Rules

Second step: rules for more complex mappings are created by combining primitive transfer rules that are adjacent in the source and target f-structures.

Example:

combining the primitive transfer rule that maps *sein* to *have* with the primitive transfer rule that maps *ich* to *I* to produce the complex transfer rule:

```
PRED(%X1,sein)              PRED(%X1,have)
 SUBJ(%X1,%X2)   ==>   SUBJ(%X1,%X2)
PRED(%X2,ich)                   PRED(%X2,I)
XCOMP(%X1,%X3)              OBJ(%X1,%X3)
```

Transfer Contiguity Constraint

1. source and target f-structures each have to be connected.
2. f-structures in the transfer source can only be aligned with f-structures in the transfer target, and vice versa.

- Analogous to constraint on contiguous and alignment-consistent phrases in phrase-based SMT
- Prevents extraction of rule that would translate *dankbar* directly into *appreciation* since *appreciation* is aligned also to *zutiefst*

PRED(%X1,dankbar)		**PRED(%X1,appr.)**
ADJ(%X1,%X2)	**==>**	**SPEC(%X1,%X2)**
in set(%X3,%X2)		**PRED(%X2,a)**
PRED(%X3,zutiefst)		**ADJ(%X1,%X3)**
		in set(%X4,%X3)
		PRED(%X4,deep)

Linguistic Filters on Transfer Rules

- Morphological stemming of PRED values
- Optional filtering of f-structure snippets based on consistency of linguistic categories
  - Extraction of snippet that translates *zutiefst dankbar* into *a deep appreciation* maps incompatible categories *adjectival* and *nominal*; valid in string-based world
  - Translation of *sein* to *have* might be discarded because of *adjectival* vs. *nominal* types of their arguments
  - Larger rule mapping *sein zutiefst dankbar* to *have a deep appreciation* is ok since *verbal* types match

Parsing &Transfer

- LFG grammars: **c(onstituent)-structures** (trees) and **f(unctional)-structures** (attribute value matrices) as output, for parsing source and target text
- FRAGMENT grammar
  - parses out-of-scope input as well-formed chunks, with unparsable tokens possibly interspersed; correct parse chosen by fewest chunk method
  - FRAGMENT grammar is used in 20% of cases

- non-deterministical application of all of the induced transfer rules in parallel
- Each fact must be transferred by exactly one rule
- Default rule transfers any fact as itself
- Transfer works on chart using parser's unification mechanism for consistency checking
- Selection of most probable transfer output is done by beam-decoding on transfer chart

Generation

- LFG grammars are used bidirectionally for parsing and generation
- Generator has to be fault-tolerant in cases where transfer-system operates on FRAGMENT parse or produces non-valid f-structures from valid input f-structures
- Robust generation from unknown (e.g., untranslated) predicates and from unknown f-structures
- Generation from unknown predicates:
  - Unknown German word “Hunde” is analyzed by German grammar to extract stem (e.g., PRED = Hund, NUM = pl) and then inflected using English default morphology (“Hunds”)
- Generation from unknown constructions:
  - Default grammar that allows any attribute to be generated in any order is mixed as suboptimal option in standard English grammar, e.g. if SUBJ cannot be generated as sentence-initial NP, it will be generated in any position as any category

Statistical Models and Training

1. Log-probability of source-to-target transfer rules, where probability *r(e|f)* or rule that transfers source snippet *f* into target snippet *e* is estimated by relative frequency

$$r(e \mid f) = \frac{count(f ==> e)}{\sum_{e'} count(f ==> e')}$$

2. Log-probability of target-to-source transfer rules, estimated by relative frequency
3. Log-probability of lexical translations *l(e|f)* from source to target snippets, estimated from Viterbi alignments *a** between source word positions *i=1, …n* and target word positions *j=1,…,m* for stems f_i and e_j in snippets *f* and *e* with relative word translation frequencies $t(e_j|f_i)$:

$$l(e \mid f) = \prod_j \frac{1}{|\{i \mid (i,j) \in a^*\}|} \sum_{(i,j) \in a^*} t(e_j \mid f_i)$$

4. Log-probability of lexical translations from target to source snippets

Statistical Models and Training cont.

5. Number of transfer rules
6. Number of transfer rules with frequency 1
7. Number of default transfer rules
8. Log-probability of strings of predicates from root to frontier of target f-structure, estimated from predicate trigrams in English f-structures
9. Number of predicates in target f-structure
10. Number of constituent movements during generations based on original order of head predicates of the constituents
11. Number of generation repairs
12. Log-probability of target string as computed by trigram language model
13. Number of words in target string

Experimental Evaluation

Experimental setup

- German-to-English translation on the Europarl parallel data set
- training and evaluation on sentences with 5 to15 words
- training set of 163,141 sentences
- development set of 1967 sentences
- test set of 1,755 sentences of length 5-15
- improved bidirectional word alignment based on GIZA++ (Och et al. 1999)

Experimental Evaluation cont.

- LFG grammars for German and English (Butt et al. 2002; Riezler et al. 2002)
- SRI trigram language model (Stocke'02)
- LFG grammars - 100% coverage on unseen data (80% parsed as full parses; 20% FRAGMENT parses)
- around 700,000 transfer rules extracted from f-structure pairs chosen according to a dependency similarity measure.
- They considered 1 German parse for each source sentence, 10 transferred f-structures for each source parse, and 1,000 generated strings for each transferred f-structure.

Experimental Evaluation cont.

- Selection of most probable translations in two steps:
  - Most probable f-structure by beam search (n=20) on transfer chart using features 1-10
  - Most probable string selected from strings generated from selected n-best f-structures using features 11-13
- Comparison with PHARAOH (Koehn et al. 2003) and IBM Model 4 (Och et al. 1999)
- To train the weights for phrase-based SMT they used the first 500 sentences of the development set
- The weights of the LFG-based translator were adjusted on the 750 sentences that were in coverage of their grammars.

Automatic Evaluation

	M4	LFG	P
in-coverage	5.13	*5.82	*5.99
full test set	*5.57	*5.62	6.40

- NIST sensitive evaluation metric & approximate randomization test for significance testing
- Experimentwise significance level of .05 achieved by reducing per-comparison significance level to .01 in 3-fold comparison (see Cohen'95)
- 44% in-coverage of grammars; 51% FRAGMENT parses and/or generation repair; 5% timeouts
  - In-coverage: Difference between LFG and P not significant
  - Suboptimal robustness techniques decrease overall quality

Manual Evaluation

- Randomly selected 500 in-coverage examples
- Two independent human judges were presented with the source sentence, and the output of the **phrase-based** and **LFG-based** systems in a blind test.
- Separate evaluation under criteria of **grammaticality/fluency** and **translational/semantic adequacy.**
- Net improvement in translational adequacy on agreed-on examples is **11.4%** on 500 sentences (**57**/500), amounting to **5%** overall improvement in hybrid system (44% of 11.4%)
- Net improvement in grammaticality on agreed-on examples is **15.4%** on 500 sentences, amounting to **6.7%** overall improvement in hybrid system

	adequacy			grammaticality		
j1\j2	P	LFG	eq	P	LFG	eq
P	**48**	8	7	**36**	2	9
LFG	10	**105**	18	6	**113**	17
equal	53	60	**192**	51	44	**223**

Discussion

- promising results for examples that are in coverage of the employed LFG grammars

HOWEVER

- high percentage of out-of-coverage examples
  - Accumulation of 2 x 20% error-rates in parsing training data
  - Errors in rule extraction
  - Together result in ill-formed transfer rules causing the generator to back-off to robustness techniques

- propagation of errors through the system also for in-coverage examples
  - Error analysis: 69% transfer errors, 10% due to parse errors

- discrepancy between NIST and manual evaluation
  - Suboptimal integration of generator, making training and translation with large n-best lists infeasible
  - Language and distortion models applied *after* generation

Conclusion

- Integration of grammar-based generator into dependency-based SMT system achieves state-of-the-art NIST and **improved grammaticality and adequacy on in-coverage examples**
- It is determinable when sentences are in coverage of system, therefore **possibility of hybrid system**

Future work

- on improvements of in-coverage translations
- on the application of the system to other language pairs and larger data sets

Thank you for your attention!

☺

Examples: LFG > PHARAOH

src: in diesem fall werde ich meine verantwortung wahrnehmen

sef: then i will exercise my responsibility

LFG: in this case i accept my responsibility

P: in this case i shall my responsibilities

src: die politische stabilität hängt ab von der besserung der lebensbedingungen

ref: political stability depends upon the improvement of living conditions

LFG: the political stability hinges on the recovery the conditions

P: the political stability is rejects the recovery of the living conditions

src: und schließlich muß dieser agentur eine kritische haltung gegenüber der kommission selbst erlaubt sein

ref: moreover the agency must be able to criticise the commission itself

LFG: and even to the commission a critical stance must finally be allowed this agency

P: finally this is a critical attitude towards the commission itself to be agency

src: nach der ratifizierung werden co2 emissionen ihren preis haben

ref: after ratification co2 emission will have a price tag

LFG: carbon dioxide emissions have its price following the ratification

P: after the ratification co2 emissions are a price

Examples: PHARAOH > LFG

src: was wir morgen beschließen werden ist letztlich material für das vermittlungsverfahren

sef: whatever we agree tomorrow will ultimately have to go into the conciliation procedure

LFG: one tomorrow we approved what is ultimately material for the conciliation procedure

P: what we decide tomorrow is ultimately material for the conciliation procedure

src: die verwaltung muß zukünftig schneller reagieren können

ref: in future the administration must be able to react more quickly

LFG: more in future the administration must be able to react

P: the administration must be able to react more quickly

src: das ist schon eine seltsame vorstellung von gleichheit

ref: a strange notion of equality

LFG: equality that is even a strange idea

P: this is already a strange idea of equality

src: frau präsidentin ich beglückwünsche herrn nicholson zu seinem ausgezeichneten bericht

ref: madam president I congratulate mr nicholson on his excellent report

LFG: madam president I congratulate mister nicholson on his report excellented

P: madam president I congratulate mr nicholson for his excellent report

References

- Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. *COLING'02, Workshop on Grammar Engineering and Evaluation*.
- Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for statistical machine translation. *MT Summit IX*.
- David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. *ACL'05*. Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. The MIT Press.
- Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. *ACL'05*.
- Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. *ACL'05*.
- George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *ARPA Workshop on Human Language Technology*.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *HLT-NAACL'03*.
- Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. User manual. Technical report, USC ISI.
- Dekang Lin. 2004. A path-based transfer model for statistical machine translation. *COLING'04*.
- Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. *Workshop on Data-Driven Machine Translation*.

- EricW. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley.
- Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. *EMNLP'99*.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *HLT-NAACL'03*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report IBM RC22176 (W0190-022).
- Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. *ACL'05*.
- Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. *ACL-05Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.
- Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. *ACL'02*.
- Stefan Riezler, Tracy H. King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. *HLT-NAACL'03*.
- Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. *International Conference on Spoken Language Processing*.
- Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. *COLING'04*.