DFKI System Combination using Syntactic Information

Christian Federmann, Sabine Hunsicker, Yu Chen, Rui Wang
DFKI Language Technology Lab

Overview

- Introduction & Motivation
- System Combination Approach
- Experiments
- Conclusion & Outlook
- Questions & Answers

Introduction & Motivation

Introduction

- We report on research conducted within the EuroMatrixPlus project
- EM+ aims at "Bringing Machine Translation for European Languages to the User"
- WP2 working on improved hybrid machine translation systems
- Work based on the Lucy RBMT system

Motivation

- Underlying assumption: different machine translation paradigms have differing strengths and weaknesses;
- often, these differences are complementary, so a clever combination of both techniques should allow to create better translations
- hence → research on hybrid MT systems

DFKI's Hybrid History

- 2009—Shallow hybrid MT system based on substitution of NPs into RBMT sentences
- 2010—Statistical System Combination and improved shallow system (more factors)
- 2011—Deeper integration by adding a stochastic parse selection component

System Combination Approach

Basic Idea

- Extending previous work on constituent substitution for hybrid MT
- One system chosen as ‘translation template’
- Remaining systems provide alternatives
- Substitution based on decision factors
- Factors are based on syntactic features

Finding the right template...

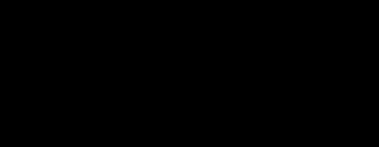
- ML4HMT shared task data contains 5 systems
- Level of annotation details varies greatly
  - Makes it difficult to equally use the data
- We decided to use Lucy RBMT as template
  - Rule-based systems create structurally sound sentences
  - Lucy provides parse tree information
  - (plus) we already worked with Lucy before...

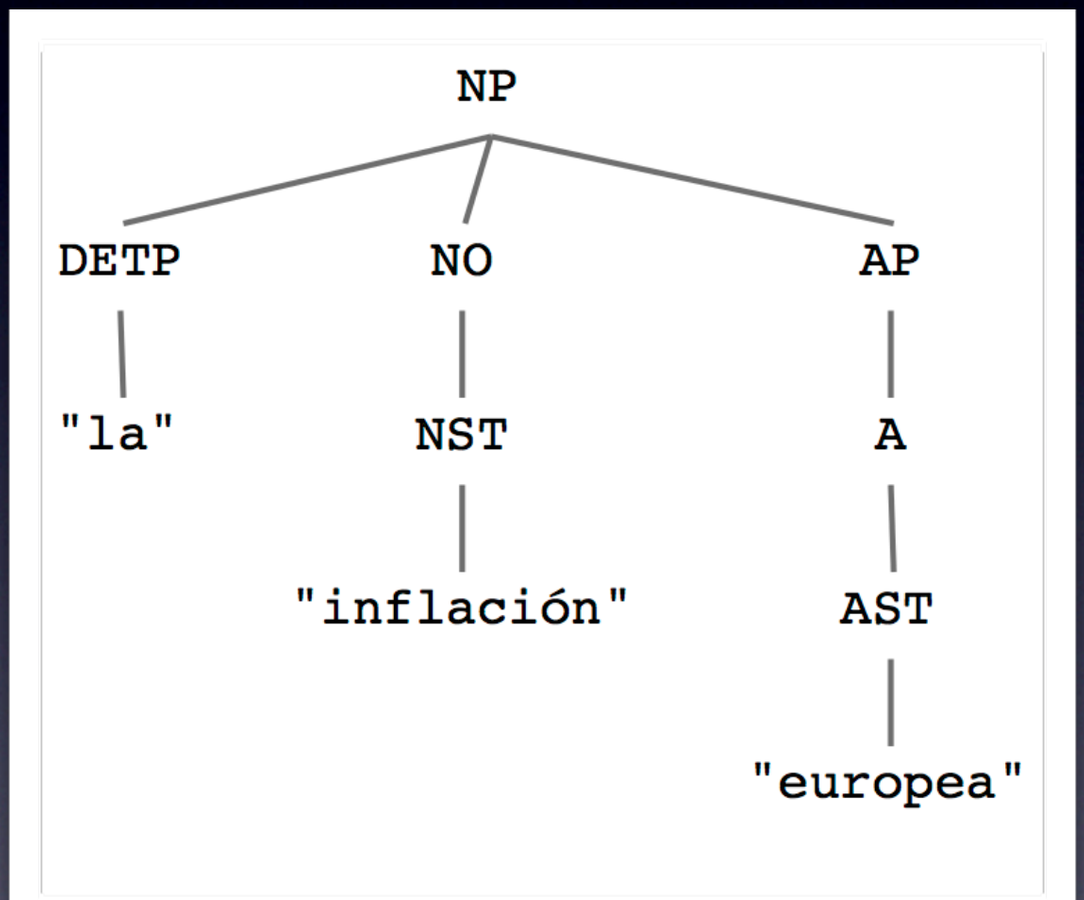
Reconstructing Parse Trees

- ML4HMT shared task data provides flattened parse trees
- We derived an algorithm to approximate the original parse trees
- Example: “la inflación europea”
- Learned heuristics regarding valid pos categories, e.g., NO can be NST or PRN

Flattened Parse Trees

```
<metanet:token id="s1_t2_r1_d1_k4">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NP"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k5">
<metanet:annotation type="alo" value="la"/>
<metanet:annotation type="can" value="el"/>
<metanet:annotation type="cat" value="DETP"/>
<metanet:string>la</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k6">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NO"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k7">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NST"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k8">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AP"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k9">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="A"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k10">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AST"/>
<metanet:string>europea</metanet:string>
</metanet:token>
```





Substitution Process

1. Compute approximated parse trees
2. Find *interesting* phrases (noun, verb, adjectives)
   - we consider noun, verb, adjective phrases
   - word alignment is computed using GIZA++
3. Each candidate translation is evaluated by some decision factors

Decision Factors

- **Matching POS?** only substitute if part-of-speech matches
- **Majority Vote** prefer more frequent translation candidates
- **Context** part-of-speech matching for left/right context
- **Language Model** fragments and left/right context scored by LM

Experiments

Overview

- Training Data
  - **Corpus:** ML4HMT shared task data
  - **Domain:** News text
  - **Size:**
    - 1,025 sentences (development)
    - 1,026 sentences (test set)
  - **Translation direction:** Spanish → English

Experimental Setup

- XML parser trained on development set
- We defined several system configurations
- Focus on comparison to Lucy baseline
  - RBMT systems usually perform bad in terms of BLEU
  - our approach cannot easily be tuned with BLEU
- Manual inspection of combination results

Feature Configurations

Configuration	Matching POS?	Context?
strict	yes	yes
pos	yes	no
context	no	yes
relaxed	no	no

Development Set

Automated Scores

Configuration	NIST	BLEU
baseline	5.5068	0.1516
strict	5.0937	0.1532
pos	5.0962	0.1534
context	**5.0984**	**0.1535**
relaxed	5.0932	0.1535

Development Set

Substitution Statistics

Configuration	# of substitutions
strict	412
pos	1,121
context	458
relaxed	1,317

Development Set

Hmmm...

Evaluation

- All combinations outperformed the baseline
- Differences in BLEU were not conclusive
- Hence, we conducted a manual evaluation
  - *context* disallows, e.g.,“it is saved” → “it is saves”
  - *context* implicitly includes part-of-speech
  - *relaxed* leads to many useless substitutions
- We finally submitted the *context* translations

Conclusion & Outlook

Conclusion

- ML4HMT shared task data allowed us to fuse translation output from different MT ‘classes’
- Single word substitution gave improvements
- Good syntactic structure of RBMT ‘skeleton’ was retained
- Lexical semantics improved by substitution

Outlook

- Investigate the contribution the different source systems have
- Extend the substitution to entire phrases and multi-word expressions
- Learn substitution rules using ML techniques
- Find ways of avoiding substitution errors
- Use parser to allow other ‘skeleton’ systems

Thank you!

Questions & Answers

Acknowledgements

This work has been funded under the Seventh Framework Programme for Research and Techno- logical Development of the European Commission through the T4ME contract (grant agreement no.: 249119).

Part of this work was supported by the EuroMatrixPlus project (IST-231720) which is also funded by the European Community.