UNIVERSITÄT DES SAARLANDES

# A Sense-Based Translation Model for Statistical Machine Translation

Deyi Xiong and Min Zhang, ACL 2014

Anastasija Amann

9 June 2018

Machine Translation 2018

Department of Language Science and Technology

# Table of contents

# Introduction

# Lexical ambiguity in Machine Translation

## Word senses for "bass" in WordNet[1]

1. *bass*: the lowest part of the musical range
2. *bass*, *bass part*: the lowest part in polyphonic music
3. *bass*, *basso*: an adult male singer with the lowest voice
4. *sea bass*, *bass*: the lean flesh of a saltwater fish of the family Serranidae
5. *freshwater bass*, *bass*: any of various North American freshwater fish with lean flesh (especially of the genus Micropterus)
6. *bass*, *bass voice*, *basso*: the lowest adult male singing voice
7. *bass*: the member with the lowest range of a family of musical instruments
8. *bass*: nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes

[1] `http://wordnetweb.princeton.edu/perl/webwn?s=bass&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=`

# Word senses in Machine Translation

*Are pure word senses useful for SMT?*

# Word senses in Machine Translation

*Are pure word senses useful for SMT?*

**Deyi Xiong and Min Zhang (2014): A Sense-Based Translation Model for Statistical Machine Translation**

1. Infer and integrate word senses into SMT system
2. Conduct experiments on Chinese-to-English translation

# Word senses in Machine Translation

*Are pure word senses useful for SMT?*

**Deyi Xiong and Min Zhang (2014): A Sense-Based Translation Model for Statistical Machine Translation**

1. Infer and integrate word senses into SMT system
2. Conduct experiments on Chinese-to-English translation

*Yes*, automatically learned word senses can improve the translation quality.

# Architecture

# Topic-based Word Sense Induction (WSI)

**Clustering problem**

- Automatically induce word senses of tokens given the surrounding contexts (= bags of k neighboring words, *pseudo documents*)

→ Distributional hypothesis: Words in the same contexts tend to have similar meanings

**Clustering algorithm**

- Predict sense clusters using topic modeling
- Hierarchical Dirichlet Process (HDP): no prespecified sense inventory or number

# Word Sense Tagger

1. Remove stop words and rare words.
2. Extract all possible pseudo documents for each source word type.
3. Train with this corpus a HDP-based WSI model for the word type. Skip highly frequent words.
4. Choose the sense with the highest probability to label the corresponding token.

# Training

- Maximum Entropy classifiers predict translation probability $p(e|C(c))$: Probability that source word $c$ is translated into target phrase $e$ given contextual information (word senses)
- Two groups of features:
  - Lexicon features: word c and its k preceding and succeeding words
  - Sense features: predicted senses in the same $\pm$k-word window
- Experiments: Sense features *do* provide new information

# Decoding

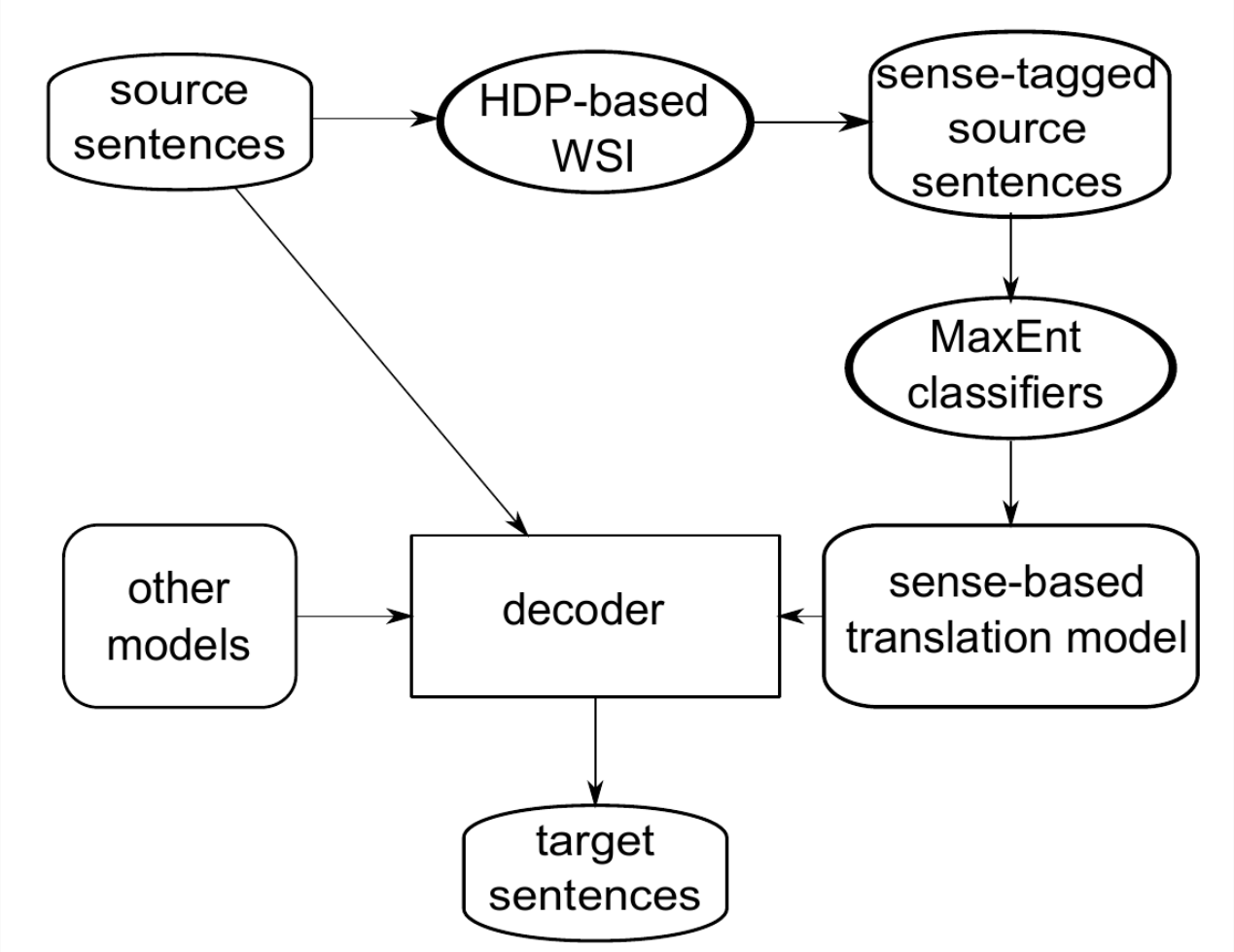



**Figure 1:** Architecture of SMT system with the sense-based translation model (Xiong and Zhang 2014).

# Experiments

*Are pure word senses useful for SMT?*

# Setup

- Baseline: state-of-the-art SMT system (Bracketing Transduction Grammars, maximum entropy based reordering model)
- Tools: Giza++, SRILM, HDP toolkit, MaxEnt tool, BLEU, NIST
- Data: 8 LDC corpora for training (Chinese-English), NIST MT03 as development set, and NIST MT05 as test set.

# WSI statistics

| | Training | Test |
|---|---|---|
| # types | 67,723 | 4,348 |
| # total pseudo documents | 27.73M | 11,777 |
| # average pseudo documents | 427.79 | 2.71 |
| # total senses | 271,770 | 24,162 |
| # average senses | 4.01 | 5.56 |

**Table 1:** Statistics of HDP-based WSI on the training and test data.

# Example word senses

| $s_1$ | $s_2$ | $s_3$ |
|---|---|---|
| 运营 (operate) | 运营 (operate) | 运营 (operate) |
| 设施 (facility) | 卫星 (satellite) | 市场 (market) |
| 计划 (plan) | 系统 (system) | 企业 (enterprise) |
| 基础 (foundation) | 国家 (country) | 竞争 (competition) |
| 项目 (project) | 提供 (supply) | 资产 (assets) |
| 公司 (company) | 国际 (inter-nation) | 利润 (profit) |
| 结构 (structure) | 机构 (institution) | 造成 (cause) |
| 服务 (service) | 进行 (proceed) | 费用 (cost) |
| 组织 (organization) | 中心 (center) | 资金 (capital) |
| 提供 (supply) | 合作 (cooperate) | 业务 (business) |
| $s_4$ | $s_5$ | $s_6$ |
| 费用 (cost) | 城市 (city) | 处于 (lie) |
| 股价 (share price) | 处理 (process) | 拍照 (photograph) |
| 27000 | 自来水 (tap-water) | 119 |
| 科索沃 (Kosovo) | 工厂 (factory) | DPRK |
| 额外 (extra) | 汽车 (car) | 保险 (insurance) |
| 工资 (wage) | 铁路 (railway) | 超支 (overspend) |
| 美元 (dollar) | 污水 (sewage) | 地位 (position) |
| 商业 (commerce) | 办事处 (office) | 经济 (economy) |
| 收入 (income) | 保本 (break-even) | 竞争者 (competitor) |
| 铁路局 (railway administration) | 部件 (component) | 平衡 (balance) |

**Figure 2:** Six different senses learned for the word "运营" from the training data.

# Are pure word senses useful for SMT?

**Table 2:** Experiment results of the sense-based translation model (STM) against the baseline.

| System | BLEU(%) |
|---|---|
| Base | 33.53 |
| SMT (sense) | 34.15 |
| SMT (sense+lexicon) | 34.73 |

## Observations

- Overall improvement of 1.2 BLEU points over the baseline
- Improvement of 0.62 BLEU points with simply word senses

# Are pure word senses useful for SMT?

**Table 2:** Experiment results of the sense-based translation model (STM) against the baseline.

| System | BLEU(%) |
|---|---|
| Base | 33.53 |
| SMT (sense) | 34.15 |
| SMT (sense+lexicon) | 34.73 |

**Observations**

- Overall improvement of 1.2 BLEU points over the baseline
- Improvement of 0.62 BLEU points with simply word senses

$\rightarrow$ *Yes*, automatically learned word senses can improve the translation quality.

# Conclusion

# Conclusion

**Summary**

- SMT can benefit from automatically inferred word senses, especially in view of lexical ambiguity.
- Word senses provide additional distributional semantic information

**Further work**

- Build and integrate a sense-based language model

$\rightarrow$ Are word senses useful for word prediction?

# Questions?

# Backup slides

# Backup slides

**Translation probability**

$$p(e|C(c)) = \frac{exp(\sum_i \theta_i h_i(e, C(c)))}{\sum_{e'} exp(\sum_i \theta_i h_i(e', C(c)))} \quad (1)$$

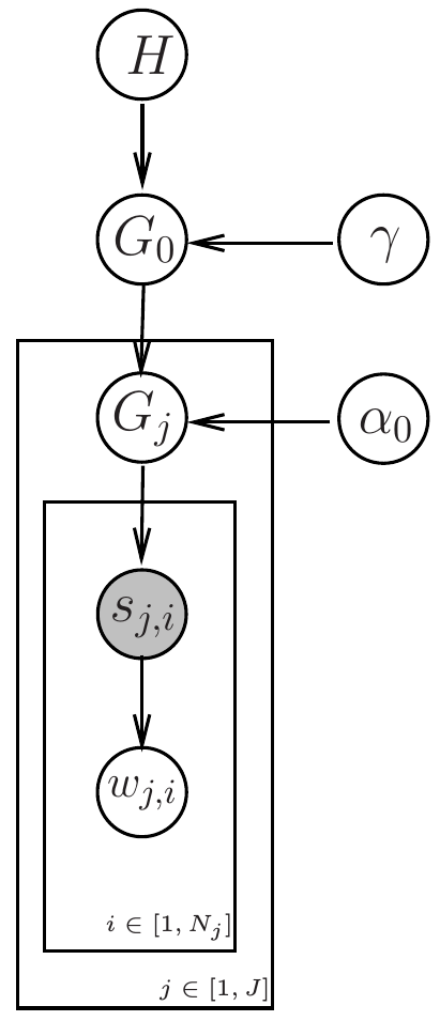
**Feature function**

$$h(e, C(c)) = \begin{cases} 1 & \text{if } e = \square \text{ and } C(c).\mu = v \\ 0 & \text{else} \end{cases}$$

- $\square$: placeholder for a possible target translation ($\leq$ 3 words or NULL)
- $\mu$: name of feature for source word $c$
- $v$: value of the feature $\mu$

## HDP generative process for WSI

1. Sample a base distribution $G_0$ from a Dirichlet process $DP(\gamma, H)$ with a concentration parameter $\gamma$ and a base distribution $H$
2. For each pseudo document $D_j$, sample a distribution $G_j \sim DP(\alpha_0, G_0)$
3. For each item $w_j, i$ in the pseudo document $D_j$,
   - 3.1. sample a sense cluster $s_{j,i} \sim G_j$; and
   - 3.2. sample a word $w_{j,i} \sim \phi_{s_{j,i}}$.



# References

Xiong, Deyi and Min Zhang. 2014. A Sense-Based Translation Model for Statistical Machine Translation, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1459–1469, Baltimore, Maryland, USA, Association for Computational Linguistics.