Joint Parsing and Alignment with Weakly Synchronized Grammars

David Burkett, John Blitzer, & Dan Klein
1) Align sentence pairs (GIZA++)
2) Parse English sentences (Berkeley parser) Parse Foreign sentences
3) Extract rules (Galley et al. 2006)
4) Tune discriminative parameters
## Data Setting for Joint Models

| English WSJ | Chinese CTB | Unlabeled parallel text | Parallel, Aligned CTB |
|-------------|-------------|-------------------------|-----------------------|
| (EN;▲)     | (中文;▲)    | (EN; 中文)              | (EN,中文;▲▲▲)        |
| (EN;▲)     | (中文;▲)    | (EN; 中文)              | (EN,中文;▲▲▲)        |
|            |             |                         | (EN,中文;▲▲▲)        |
|            |             |                         | (EN,中文;▲▲▲)        |
| (EN;▲)     | (中文;▲)    | (EN; 中文)              | (EN,中文;▲▲▲)        |
| 在 | 办公室 | 里 | 读了 | 书 |
|---|---|---|---|---|
| at | office | in | read | book |

read
the
book
in
the
office
Build a model $p_\theta (\text{△}, \text{△}, \text{中文, EN})$
Correspondence via Synchronous Grammars
Synchronous derivation
Synchronous Derivation
Weakly Synchronized Example

... were established in such places as Quanzhou, Zhangzhou, etc.

在泉州, 漳州等地设立了...
Weakly Synchronized Example

Separate PCFGs
... were established in such places as Quanzhou, Zhangzhou, etc.
Weakly Synchronized Example

Points for synchronization, but not required

在 泉州 漳州 等 地 设立 了 ...
Correspondence Model & Feature Types

Feature type 1: Word Alignment
\[ \phi([-\text{中文} - \text{EN}], \text{中文}) \]  
office

Feature type 2: Monolingual Parser
\[ \phi([-\text{中文} - \text{EN}], \text{中文}) \]  
in the office

Feature type 3: Correspondence
\[ \phi([-\triangle, \triangle, \text{中文}]) \]
Estimating $\theta$

- Set $\theta$ to maximize the log-likelihood of the correct parses & alignments

$$\log p_\theta (\triangle, \bigtriangleup, \text{EN,中文}) = \langle \theta, \phi(\triangle, \bigtriangleup, \text{EN,中文}) \rangle - \log Z(\text{EN,中文})$$

- $Z(\text{EN,中文})$ normalizes $p_\theta$ to sum to 1

$$Z(\text{EN,中文}) = \sum_{\triangle, \bigtriangleup, \text{EN,中文}} \exp \{ \langle \theta, \phi(\triangle, \bigtriangleup, \text{EN,中文}) \rangle \}$$
Computing \( Z(EN, \text{中文}) \)

\[
Z(EN, \text{中文}) = \sum \exp \{ \langle \theta, \phi(\triangle, \triangle, \square, \text{EN,中文}) \rangle \}
\]

Individual \( \sum \), \( \sum \), \( \sum \) have polynomial-time dynamic programming algorithms

Correspondence features tie pieces together

Computing \( Z(EN, \text{中文}) \) exactly is intractable
Approximating $Z^{(EN,中文)}$: Mean Field

- Exploit tractability in individual models: $\sum_{\Delta} \sum_{\Delta} \sum_{\Xi}$

- Factored approximation: $p_\theta(\Delta, \Delta, \Xi | EN,中文) \approx q(\Delta)q(\Delta)q(\Xi)$

- Set $q$ to minimize $KL(q(\Delta)q(\Delta)q(\Xi), p_\theta(\Delta, \Delta, \Xi | EN,中文))$

**Algorithm**

1) Initialize $q(\Delta)q(\Delta)q(\Xi)$ separately

2) Iterate:

$$q(\Delta) \propto \exp \{ \langle \theta, \phi(\Delta, E_q(\Delta), E_q(\Xi)) \rangle \}$$

$$q(\Delta) \propto \exp \{ \langle \theta, \phi(E_q(\Delta), \Delta, E_q(\Xi)) \rangle \}$$

$$q(\Xi) \propto \exp \{ \langle \theta, \phi(E_q(\Delta), E_q(\Delta), \Xi) \rangle \}$$
Large scale inference

We can approximate $Z(\text{EN}, \text{中文})$ in polynomial time, but . . .

$$\sum \text{Sum over possible alignments is an } O(n^6) \text{ algorithm.}$$

But computers are fast, right?

- Medium-length sentences are 50 words long
- Small translation data sets are 250,000 sentences
- ~4 quadrillion operations (See [BBK10, HBDK09] for speedup details)
Quantitative Results: Parsing

- Monolingual
- Joint
Quantitative Results: Parsing

|                | Monolingual | Joint  |
|----------------|-------------|--------|
| Chinese parser | 83.6%       | 85.7%  |
Quantitative Results: Parsing

|       | Monolingual | Joint  |
|-------|-------------|--------|
| Chinese parser | 81.2%       | 84.5%  |
| English parser  | 81.2%       | 84.5%  |
Incorrect English PP Attachment
Corrected English PP Attachment

... were established in such places as Quanzhou, Zhangzhou, etc.

在泉州、漳州等地设立了...
Quantitative Results: Translation

BLEU improvement from 29.4 to 30.6
At this point the cause of the plane collision is still unclear. The localCAA will launch an investigation into this.

**Baseline (GIZA++):**
The cause of planes is still not clear yet, local civil aviation department will investigate this.

**Bilingual Adaptation Model:**
The cause of plane collision remained unclear, local civil aviation departments will launch an investigation.
Thanks