cross-lingual tagging

YAROWSKY & NGAI, 2001
CUCERZAN & YAROWSKY, 2002

LING 575 • 19 FEB 2008
Ankit Kumar Srivastava
UNSUPERVISED TAGGING

- Filtered Lexicon
- Clustering and Prototypes
- Cross-lingual Projection
- ...
CROSS-LINGUAL APPROACH

- Source Language (SL) Tagger
- Projection & Induction: SL → TL
- Tagging Target Language (TL)

Leveraging unannotated parallel text of resource-rich and resource-poor languages
YAROWSKY & NGAI ‘01

Inducing Multilingual POS Taggers and NP Brackets via Robust Projection across Aligned Corpora

David Yarowsky† and Grace Ngai‡‡
{yarowsky,gyn}@cs.jhu.edu

† Johns Hopkins University ‡‡ Weniwen Technologies
Baltimore, MD 21218, USA Hong Kong

- Algorithm
- Experiments
- Discussion

Induce a French POS tagger using automatically word-aligned Canadian Hansards text, tagged on the English side

Ling 575: 19 February, 2008
Cross-lingual Approach: 4

Skip sxn 5 (np brackets)
YN01: ALGORITHM

- EOS detection, tokenization
- Word Alignment (IBM Model 3)
- Tag SL (Brill 1995)
- Direct POS Projection (1-1, 1-n)
- Bigram Tagger Induction
- Evaluation
YN01: ALGORITHM

- After projection, build and re-estimate bigram tagger for TL, with lexical prior estimation and tag sequence model estimation (Noise Robust Bigram Induction)

\[
\arg\max_T P(T|W) = P(W|T)P(T)
\]

\[
P(W|T) = P(w_1...w_n|t_1...t_n) \approx \prod_{i=1} P(w_i|t_i)
\]

\[
P(T) = P(t_1...t_n) \approx P(t_1)P(t_2|t_1)P(t_n|t_{n-1})
\]
YN01: ALGORITHM

- Lexical Prior Estimation

\[
P(w_i | t_i) = \frac{P(t_i | w_i)P(w_i)}{\sum_j P(t_i | w_j)P(w_j)}
\]

\[
\hat{P}(t_{(2)} | w) = \lambda_1 P(t_{(2)} | w) \quad \text{where } \lambda_1 < 1.0
\]

\[
\hat{P}(t_{(1)} | w) = 1 - \hat{P}(t_{(2)} | w)
\]

\[
\hat{P}(t_{(c)} | w) = 0 \quad \text{for all } c > 2
\]

\[
P(t | w) = \lambda_2 P_{1-to-1}(t | w) + (1 - \lambda_2) P_{1-to-n}(t | w)
\]
YN01: ALGORITHM

- Tag Sequence Model Estimation is executed separately from Lexical Priors so that high-quality data can be filtered and used
- Identifying high-quality data
- Pseudo-divergence weighting

\[ P(T) = P(t_1 \ldots t_n) \approx P(t_1)P(t_2|t_1) \ldots P(t_n|t_{n-1}) \]

\[ \frac{1}{k} \sum_{i=1}^{k} \log \hat{P}(\text{projected-tag}_i|w_i) \]
### YN01: EXPERIMENTS

- Parallel Text from Canadian Hansards
- (d) trained on 500K, (e) trained on 100K
- Evaluation on 1000-word segment (left), 120K(right)

<table>
<thead>
<tr>
<th>Model</th>
<th>Evaluate on E-F Aligned French</th>
<th>Evaluate on Unseen Monolingual French</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Core Tagset</td>
<td>Eng Eqv Tagset</td>
</tr>
<tr>
<td>(a) Direct transfer (auto-aligned, auto-project)</td>
<td>.76</td>
<td>.69</td>
</tr>
<tr>
<td>(b) Direct transfer (hand-aligned, auto-project)</td>
<td>.85</td>
<td>.78</td>
</tr>
<tr>
<td>(c) Standard bigram model (auto-aligned, auto-project)</td>
<td>.86</td>
<td>.82</td>
</tr>
<tr>
<td>(d) Noise-robust bigram induction (auto-aligned, auto-project)</td>
<td>.96</td>
<td>.93</td>
</tr>
<tr>
<td>(e) Standard bigram model (trained on heldout goldstandard)</td>
<td>.97</td>
<td>.96</td>
</tr>
</tbody>
</table>

**Inducing Multilingual POS Taggers and NP Brackets via Robust Projection across Aligned Corpora**
YN01: DISCUSSION

- Core tags
  - Does aggressive re-estimation bias help?
- Sparse Model
  - Separate Tag induction from lexical induction
- Annotation Granularity Issue
  - TL tagset dependent on SL tagset
- Poor alignment accuracy
  - Better alignment strategies since 2001?
  - Feedback POS into improving alignment

Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora
Inducing a Romanian and Spanish tagger using multitude resources but no training data

CUCERZAN & YAROWSKY ‘02

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

Silviu Cucerzan and David Yarowsky
Department of Computer Science and Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218 USA
{silviu,yarowsky}@cs.jhu.edu

- Algorithm
- Experiments
- Discussion
CY02: ALGORITHM

- SL Tagged Text (Brown, WSJ)
- Bilingual dictionary
- Morphological Analysis
- Trigram Tagger Induction
- Evaluation
CY02: ALGORITHM

- Unweighted mixture of the prior POS distributions for English word

![Table and formula showing POS distributions and probabilities.]

**Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day**
Handling multiple POS for 1 English word and phrasal English translations (independence assumption)

Use a single foreign language – English dictionary to train probabilities on English words

\[
P(T_f | w_{e1} w_{e2}) = \sum_{T_{e1}} \sum_{T_{e2}} P(T_f | T_{e1} T_{e2}) \cdot P(T_{e1} | w_{e1}) \cdot P(T_{e2} | w_{e2})
\]
CY02: EXPERIMENTS

<table>
<thead>
<tr>
<th>Target Language</th>
<th>Training Dictionary</th>
<th>Accuracy Exact POS</th>
<th>Correct POS Over Threshold</th>
<th>Coverage</th>
<th>Mean Probability of Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romanian</td>
<td>Spanish - English</td>
<td>92.9</td>
<td>97.8</td>
<td>98</td>
<td>.91</td>
</tr>
<tr>
<td>Kurdish</td>
<td>Spanish - English</td>
<td>76.8</td>
<td>93.1</td>
<td>95</td>
<td>.82</td>
</tr>
<tr>
<td>Spanish</td>
<td>Romanian - English</td>
<td>83.3</td>
<td>94.9</td>
<td>97</td>
<td>.86</td>
</tr>
</tbody>
</table>

- Romanian, Kurdish, Spanish taggers from bilingual (-EN) dictionary
Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

Rules for inflection and associated POS from Grammar reference book

<table>
<thead>
<tr>
<th>Root Affix</th>
<th>Inflected Affix</th>
<th>Part-of-speech Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o$</td>
<td>o$</td>
<td>Adj-masc-sing</td>
</tr>
<tr>
<td>o$</td>
<td>os$</td>
<td>Adj-masc-plur</td>
</tr>
<tr>
<td>o$</td>
<td>a$</td>
<td>Adj-fem-sing</td>
</tr>
<tr>
<td>o$</td>
<td>as$</td>
<td>Adj-fem-plur</td>
</tr>
<tr>
<td>e$</td>
<td>e$</td>
<td>Adj-masc,fem-sing</td>
</tr>
<tr>
<td>e$</td>
<td>es$</td>
<td>Adj-masc,fem-plur</td>
</tr>
<tr>
<td>ar$</td>
<td>o$</td>
<td>Verb-Indic Pres-p1-sing</td>
</tr>
<tr>
<td>ar$</td>
<td>as$</td>
<td>Verb-Indic Pres-p2-sing</td>
</tr>
<tr>
<td>ar$</td>
<td>a$</td>
<td>Verb-Indic Pres-p3-sing</td>
</tr>
<tr>
<td>ar$</td>
<td>amos$</td>
<td>Verb-Indic Pres-p1-plur</td>
</tr>
<tr>
<td>ar$</td>
<td>áis$</td>
<td>Verb-Indic Pres-p2-plur</td>
</tr>
<tr>
<td>ar$</td>
<td>an$</td>
<td>Verb-Indic Pres-p3-plur</td>
</tr>
</tbody>
</table>
CY02: ALGORITHM

- Hypothesize paradigms for TL words [LEFT]
- Verify with weighted Levenshtein alignment (from monolingual TL corpus) [RIGHT]
- Supervised data entry for closed class words

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day
Model re-estimation

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

CY02: ALGORITHM

- A suffix-based part-of-speech probability model $P(pos_j|\text{suffix}(w_i))$ using hierarchically smoothed tries is trained on the raw initial tag distributions, yielding coverage to unseen words and smoothing of low-confidence initial tag assignments.

- Paradigmatic cross-context tag modeling is performed as in Cucerzan and Yarowsky (2000) when sufficiently large unannotated corpora are available.

- Sub-part-of-speech contextual agreement for features such as gender is performed as described in Section 4.1.

- The part-of-speech tag sequence models $P(pos_2|pos_1, pos_0)$ utilize a weighted backoff between fine-grained and coarse-grained tags.

- Both the tag-sequence and lexical prior models are iteratively retrained using these additional evidence sources and first-pass probability distributions.
CY02: ALGORITHM

- Agreement models for gender-marked words, etc. within a window of +/- 1

\[
P(\text{Gen}_k|w) = \frac{1}{N} \sum_{i \in \text{loc}(w)} \sum_{j=-3}^{+3} P(\text{Gen}_k|w_{i+j})W_t(j)
\]
CY02: EXPERIMENTS

- Bootstrapping on 123K words (Romanian) and 3.2M words (Spanish)
- Seed dictionaries 42K (Spanish) and 7K (Romanian)
- Test data of 1000 words

**Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day**

<table>
<thead>
<tr>
<th></th>
<th>Spanish NNS 8h</th>
<th>Romanian NNS 8h</th>
<th>NNS-8h NS-4h</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All words</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core-tag</td>
<td>93.1</td>
<td>86.3</td>
<td>89.2</td>
</tr>
<tr>
<td>exact-match</td>
<td>86.5</td>
<td>68.6</td>
<td>75.5</td>
</tr>
<tr>
<td>exact w/o gender</td>
<td>87.0</td>
<td>76.7</td>
<td>83.0</td>
</tr>
<tr>
<td><strong>Nouns</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core-tag</td>
<td>90.3</td>
<td>97.4</td>
<td>97.4</td>
</tr>
<tr>
<td>*number</td>
<td>100.0</td>
<td>97.4</td>
<td>98.9</td>
</tr>
<tr>
<td>*gender</td>
<td>100.0</td>
<td>54.9</td>
<td>64.7</td>
</tr>
<tr>
<td>*definiteness</td>
<td>–</td>
<td>96.6</td>
<td>93.7</td>
</tr>
<tr>
<td>*case</td>
<td>–</td>
<td>97.4</td>
<td>97.4</td>
</tr>
<tr>
<td><strong>Verbs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core-tag</td>
<td>94.7</td>
<td>87.9</td>
<td>89.5</td>
</tr>
<tr>
<td>*tense</td>
<td>93.0</td>
<td>92.6</td>
<td>93.2</td>
</tr>
<tr>
<td>*number</td>
<td>100.0</td>
<td>91.5</td>
<td>91.2</td>
</tr>
<tr>
<td>*person</td>
<td>97.2</td>
<td>92.6</td>
<td>93.2</td>
</tr>
<tr>
<td><strong>Adjectives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core-tag</td>
<td>79.7</td>
<td>78.6</td>
<td>81.5</td>
</tr>
<tr>
<td>*gender</td>
<td>100.0</td>
<td>81.3</td>
<td>82.2</td>
</tr>
<tr>
<td>*number</td>
<td>100.0</td>
<td>98.3</td>
<td>98.3</td>
</tr>
</tbody>
</table>
MWEs from dictionaries
- First-pass on only single words

Is such a supervised effort justified?
- End justifies the means?

Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day