Induction of named-entity tagger

Continued from Yarowsky et al [2001]...

- Named entities: person names, place names, organization names, ...
- Word-based classification:
  - for each foreign word, compute probability that it is a NE
- Sum over all English words aligned to foreign word

\[
P(\text{NEClass} | f) = \sum_{e} P(\text{NEClass} | e_{i})P(e_{i} | f)
\]

- Obtain initial probability distribution from (noisy) projection through alignment
- Train NE learner (co-training algorithm) on projected output

Induction of morphological analyzers

- Goal: infer relationship between roots and inflected forms in foreign language, e.g. croire-croyaient
- Infer by summing over all English words aligned with inflected form:

\[
P(f_{out} | f_{inf}) = \sum_{i} P(f_{out} | e_{i})P(e_{i} | f_{inf})
\]

- Example:

\[
P(\text{croyaient}) = P(\text{croyaient}) + P(\text{croyaient}) + P(\text{croyaient})
\]

- Restrict to highest-confidence alignments

Induction of named-entity tagger

- Improving noisy alignments:
  - confidence filtering: discard alignments whose scores is below confidence threshold
  - POS filtering: candidate NE words have to be proper nouns
- Performance improved from initial 64% accuracy to 85% on English-French data

Induction of morphological analyzers

- Generalization to unseen words (e.g. what is P(ire|yaient)?)
  - Cluster words with similar surface form changes
  - Train supervised classifier
- Evaluation
  - English-French: 99.4% precision, 99.9% coverage
  - English-Czech: 87.8% precision, 99.1% coverage
  - English-Spanish: 96.6% precision, 98.5% coverage

In this lecture

- More on cross-lingual projection
  - NE tagging
  - Morphological analysis
  - Parsing
  - Semantic representations
- Automatic Summarization
  - Approaches
  - Key studies
  - Evaluation measures
Induction of morphological analyzers

- Performance correlated with size of training data
- Even small amount of training data (120K) can yield high performance > 90%
- Experiments using Bible as training text and Hansard data as test demonstrate cross-genre generalization capabilities
- Multiple input texts can be used (e.g., different versions of Bible)
- Bridging to third language (e.g., English to Spanish via French) works as well as using multiple input texts

Induction of parsers

- Collection of training data for parsers (treebank annotations) takes several person-years
- Projection of shallow grammatical information across languages works but what about more complex syntactic relationships?
- Hwa et al. [2004]: investigate cross-linguistic correspondence assumption for dependency-based representations:
  
  Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures TreeE and TreeF, if Nodes xE and yE of TreeE are aligned with nodes xF and yF of TreeF, respectively, and if syntactic relationship R(xE,yE) holds in TreeE, then R(xF,yF) holds in TreeF.

- Homomorphism of dependency graphs

Induction of parsers

- Example of learned morphology:

Induction of parsers

- Example

  English: I got a gift for my brother.
  Basque: Nik nire anaiari opari bat erosi nion.

  Dependency relationships:
  
  English  Basque
  Verb-subject got - I erosi - nik
  Verb-object got - gift erosi - opari
  Noun-det gift - a opari - bat
  Noun-mod brother - my anaiari - nire

Induction of parsers

- Example

  Direct project of English dependencies onto Spanish and Chinese parallel data (100 sentences) with manually generated trees and word alignments:
  - 37% F-score on Spanish
  - 38% F-score on Chinese
**Induction of parsers**

- Problem: language-specific knowledge is needed
  - Differences in morphology/realization of grammatical information, e.g.
    - English has no aspectual markers \(\Rightarrow\) Chinese aspectual markers will be left unattached
    - Spanish: Va a dormirse - He/she will fall asleep: no attachment of 'se'
  - Hand-written post-projection correction rules for handling of closed-class categories
    - Improvement to 70.3% F-score for Spanish and 67.3% F-score for Chinese
    - Sufficiently high for training new parser

- Experiments on inducing Spanish parser from English under realistic conditions:
  - Parser trained on English WSJ is applied to English side of parallel text (bible, FBIS, UN text)
  - IBM model 4 used for word alignment
  - Direct projection, followed by post-correction
  - Filtering to prune out low-quality projected trees
  - Discard if more than N% of English words have no counterpart
  - Discard if more than N% of Spanish words have no counterpart
  - Discard if more than N Spanish words are aligned to same English word
  - Train new dependency parser on foreign language

**Induction of parsers**

- Evaluation:
  - Manual vs automatic input: decrease in F-score from 70.3% to 65.7%
  - Final trained parser w/ filtering: 72.1%
  - Commercial parser: 69.2%

- English and Spanish are quite similar, what about more different languages?
  - Repeat experiment with Chinese:
    - Higher word alignment error: 41% compared to 24.4% in Spanish
    - Degrades more under automatic parsing and word alignment: 67.3% F-score (manual) vs. 52.4% (automatic)
    - Best performance with final parser + filtering: 64.3%

**Induction of semantic roles**

- Cross-lingual projection possible at even higher level?
- Shallow semantic parsing: identify semantic roles associated with syntactic constituents
- Relevant for e.g. information extraction, question answering,
- Statistical approaches to semantic parsing relatively recent (due to lack of data)
- Which representation?
  - Pado & Lapata (2005): Frame semantics

**Induction of semantic roles**

- Pilot study on English and German:
  - Gold standard annotation of parallel corpus (Europarl)
  - Sentence pair selection:
    - frames must be represented in English FrameNet
    - Sentences must have at least 1 aligned word and 1 frame in common (checked against English and German FrameNets)
  - Random selection of subset of 1,140 sentences
  - Annotation of frames and semantic role assignment by 2 bilingual speakers on single side only

---

**Frame:** describes situation

**Frame elements (FEs):** are semantic roles

**Frame-evoking elements (FEEs):** lexico-syntactic realization of FEs

FrameNet databases under development for English, German, Spanish, Japanese

- Largest annotation project: English FrameNet at Berkeley [Fillmore 2003]
- 513 frames (7,125 lexical items) in 6 years

Can annotation be accelerated by cross-lingual projection?
Induction of semantic roles

• Inter-annotator agreement:

<table>
<thead>
<tr>
<th>Measure</th>
<th>English</th>
<th>German</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame Match</td>
<td>0.99</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>Role Match</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Span Match</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.86</td>
<td>0.90</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Frame match: ratio of common frames
Role match: ratio of common roles
Span match: ratio of roles with identical spans
Kappa: standard measure of agreement between human annotators:

\[ Kappa = \frac{\text{observed agreement} - \text{chance agreement}}{\text{total agreement} - \text{chance agreement}} \]

Induction of semantic roles

• Degree of match across languages:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame Match</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Role Match</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
</tr>
</tbody>
</table>

(Using German as gold standard)

• Role agreement is high \( \Rightarrow \) when frames match across languages and argument structures are not too different, projection should be possible
• Frame matching involves more ambiguities

Induction of semantic roles

• Word-based projection:
  - project each role source role \( r \) with span \( s(r) \) to the set of all target tokens aligned to a token in \( s(r) \)
  - Semantic roles mostly span contiguous segments
  - Problem: non-contiguous projections
  - Heuristic fix: include all elements between first and last target words aligned to source span

He asked John and Mary, too.

Er fragte Johann und auch Maria.

Induction of semantic roles

• Constituent-based projection:
  - Handle word alignment errors by projecting from parse trees: align similar constituents
  - Similarity function for constituents \( c_s, c_t \):
    - \( \text{align} (c_s, c_t, \text{sim}) = \arg \max_{c_{t}} \prod_{c_{s} \in c_{t}} \text{sim}(c_s, c_t) \)
  - For each source constituent \( c_s \) in the span of a role \( r \), find most similar constituent on target side:
    - Can also be done in reverse direction (backward model)

Induction of semantic roles

• Experimental results on role assignment (comparison against human annotations)
  - Word-based projection model
    - No use of linguistic information (e.g. POS)

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>0.41</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>cw</td>
<td>0.46</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>Upper bound</td>
<td>0.35</td>
<td>0.84</td>
<td>0.84</td>
</tr>
</tbody>
</table>

with heuristics

Induction of semantic roles

• Result for constituent-based model

<table>
<thead>
<tr>
<th>Model</th>
<th>0-skip</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>0.70</td>
<td>0.33</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>cw</td>
<td>0.66</td>
<td>0.32</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Upper bound</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td></td>
</tr>
</tbody>
</table>

• Fc: forward, bc: backward, o: overlap measure only defined on content words, 0: skip unaligned constituents
• Problem: word alignment gaps
Induction of predicate-argument structure

- Manual inspection of English FEEs and German translations for two frames
- Verb FEEs in English (increase) often translated by adjectives in German (höher) ⇒ different semantic roles
- Analysis of 122 sentences from parallel English-German Europarl corpus: are the same frames evoked by translation pair

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPPOS</td>
<td>CPPOS</td>
<td>73</td>
</tr>
<tr>
<td>CCOSP</td>
<td>CCOSP</td>
<td>49</td>
</tr>
</tbody>
</table>

- Often frame mismatch ⇒ single frames not suitable as units for cross-linguistic matching

Induction of predicate-argument structure

- Frames are often embedded within other frames; embedded structure matches single frame in other language:
- Define frame groups as units for projection
- Not possible to apply projection to each base frame + embedding frame ⇒ overgeneration

Induction of predicate-argument structure

- Constrain possible frame groups:
  - Semantic head of embedding role is either FEE of base or base FEE is modifier of head
  - Example:
    - though world trade can of course increase prosperity
    - der Welthandel einen höheren Wohlstand zur Folge hat

Two frame groups are aligned if the FEEs of their base frames are aligned

Induction of predicate-argument structure

- Roles are projected via word alignments of semantic head words:
  - CAUSE ITEM
  - neutral: [through world trade can of course increase prosperity]
  - neutral: [der Welthandel einen höheren Wohlstand zur Folge hat]

Induction of predicate-argument structure

- Algorithm:
  - Identify target frame group
  - Identify partial paraphrases (base frames)
  - Iteratively extend base frames, linking roles, until all semantic roles of target frame group are filled
  - Iteration:
    - Find embedding frames on one side only (input language side)
    - Stop when no extensions can be found

- Does algorithm correctly match roles across languages?
  - Successful in about 61% of all analyzed cases
- Problems:
  - languages may realize role structures differently (eg German emphasizes event, English emphasizes action)
  - Algorithm may need to be extended to handle larger frame groups
Conclusions

- Cross-linguistic data project easier for lower levels of language (shallow syntactic information, morphology)
- Accuracy decreases for higher levels (syntactic relations/dependencies, semantic role information)
- Projection at higher levels requires more monolingual information, more hand-crafted resources (e.g., ontologies, semantic networks,...)

Automatic Summarization

- Goal: distill most important information from a document and present it in compact form
  - Single document vs. multiple documents (multidocument summarization)
  - From text, speech, multimedia...
  - From documents in a single language vs. documents in multiple languages
  - Query-based vs. generic summarization
  - Indicative: indicates type of information
  - Informative: quantitative/qualitative information
  - Evaluative: critical assessment of content

Automatic summarization: basic terms

- Extractive summarization: summarization by extracting key sentences/phrases/words from documents, not necessarily coherent

- Abstractive summarization: distill information and present it in coherent form, generation process involved

Automatic summarization: demo

- Multidocument extractive summarization of news texts (online demo):
  - [http://lada.si.umich.edu:8080/clair/nie1/nie.cgi](http://lada.si.umich.edu:8080/clair/nie1/nie.cgi)
Automatic summarization: approaches

- Early approaches:
  - Luhn (1958): term frequencies used to score sentences according to relevance
    - Filters out frequent function words
    - Clusters words based on orthographic similarity
    - Stop-word lists and stemming are common today
  - Edmundson (1969): structural information
    - Position of sentence within paragraph/document
    - Lexical cues: “this indicates that…”, “a significant result”, “our conclusion is…”

- AI approaches (1980’s)
  - Rely on knowledge representation formalisms
    - E.g. propositional logic, scripts, basic linear representation, ...
  - Build up representation for individual sentence unit
  - Augment by representation of discourse structure
  - Assess importance of sentence units
  - Generate natural language summary
  - Examples: FRUMP [deJong 1982], SUSY [Fum et al. 1985]

Statistical approaches

- Since 1990s: statistical approaches
  - Emphasis on trainability: adaptation to particular types of text
  - Classification approach: classify parts of input documents as to whether they should be part of summary
  - Generation approach: statistical model mimicking human abstracting process

Classification approaches

- Kupiec [1995]:
  - Bayesian classifier to compute probability of including sentence in summary given features of input sentence
  
  \[
  P(s \in E | f_1, ..., f_p) = \frac{P(s \in E)P(f_1, ..., f_p | s \in E)}{P(f_1, ..., f_p)}
  \]

  - Features: sentence length, presence of cue word, position within paragraph, thematic words, casing, ...
  - Can be supplemented by reranking

Including lexical connectedness

- Simple extraction does not take discourse structure into account:
  - Which paragraphs are connected?
  - Anaphora, coreference

- [Barzilay & Elhadad 1999]: lexical chains
  - Lexical chains: words related by collocation, repetition, synonymy, hyponymy

  Mr. Kenny is the person that invented an anesthetic machine which uses micro-computers to control the rate at which an anesthetic is pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anesthetic into the patient.
Including lexical connectedness

• Identify chains:
  - Select set of candidate words (e.g., all words appearing as nouns in WordNet)
  - Find appropriate chain based on relatedness criteria
    • Distance of positions in text
    • Distance in WordNet database
    • 3 relations strengths: extra strong > strong > medium strong
  - Insert word into chain with appropriate sense and update other word senses; create new chain if no chain is found

• Score chains: assign product of length and homogeneity index as score

\[ \text{Score(chain)} = \frac{\sum \text{count}(\text{types}_i)}{\sum \text{count}(\text{tokens}_i)} \times (1 - \frac{1}{\sum \text{count}(\text{tokens}_i)}) \]

Including lexical connectedness

• Use chains for extraction:
  - Select chains with high scores
  - For each chain, choose the sentence from the source document that contains the first appearance of a (representative) chain member
  - Representative words: words with frequency no less than average frequency of words in chain

Generation/abstractive approaches

• Top-down procedure:
  - Define set of information types to include in summary (information templates)
  - Fill slots with information from document(s)
  - Present information in coherent form (generation)

Generation/abstractive approaches

• Cut-and-paste based summarization (Jing & McKeown 2000):
  - Extract key sentence from original document
  - Generate summary by editing them
  - Identify editing operations from comparison of source documents and human abstracts
    • 1) sentence reduction: delete parts of sentences
    • 2) sentence combination: merge material from several sentences
    • 3) syntactic transformation: e.g., change voice
    • 4) lexical paraphrasing: replace words/phrases with synonyms
    • 5) generalization/specification: replace phrases with more general/more specific descriptions
    • 6) reordering: change order of sentences within an abstract
Generation/abstractive approaches

- Probabilistic model for edit operations:
  - Is sentence in summary constructed by cutting/pasting elements from source document?
  - If so, which word in summary comes from which word in source document?
- Probabilistic generative model: HMM
  - Underlying sequence of states = words in source document
  - Observations = words in summary
- Constrained by heuristic rules:
  - Adjacent words usually generate adjacent words

- Probability of removal/reduction further constrained by
  - Grammar: only delete grammatically optional constituents (based on parser/lexicon information)
  - Context: check word's relatedness to other words in context and assign importance score
- Sentence combination:
  - Decide when to combine fragments based on manual rules

Generation/abstractive approaches

- [Knight & Marcu 2000]: noisy channel model
  - Parse document and construct a large parse tree
  - Then hypothesize various smaller trees and pick the hypothesized tree with the maximum probability

\[ \text{arg max} \ P(t_x | t_y) \ P(t_y) \]

- Probabilities are computed over trees, not strings
- Parse document and construct a large parse tree
- Then hypothesize various smaller trees and pick the hypothesized tree with the maximum probability

\[ \text{arg max} \ P(t_x | t_y) \ P(t_y) \]

Generative/abstractive approaches

- Templatic rules for expanding trees:

\[
P(G \rightarrow H \ A | G \rightarrow H \ A) \ P(A \rightarrow C \ B \ D | A \rightarrow C \ D) \ P(B \rightarrow Q \ R | B) \ P(Q \rightarrow Z | Q) \ P(Z \rightarrow c | Z) \ P(R \rightarrow d | R) \]

- Rules and their probabilities are learned from parallel corpus of source documents and abstracts
- Parse both sides and identify corresponding syntactic nodes
- Decoding: for each possible compression length, extract highest-scoring trees
- Probability of reduced tree: combination of parse-tree score and bigram probability
- Note: only one edit operation (reduction) modeled