Decoder-Guided Backoff

Using Word Lattices to Improve Translation from Morphologically Complex Languages

Chris Dyer
University of Maryland
Outline this talk

- What is morphology and why does it matter to MT?
- Prior work
- Modeling morphology as observational ambiguity
- Decoding word lattices
- Experimental results
What is morphology? A crash course in words

- An important observation: words have complex internal structure.

![Cat illustration]
What is morphology?
A crash course in words

- An important observation: words have complex internal structure.
Morphology

- Conventional division:
  - Derivational morphology
    - “Derive” new forms from a root
    - Adjective $\rightarrow$ Verb (wide $\rightarrow$ widen)
    - Verb $\rightarrow$ Noun (destroy $\rightarrow$ destruction)
  - Inflectional morphology
    - “Add meaning” to a base category
    - $+\text{PLURAL}$ (cat $\rightarrow$ cats)
    - $+\text{DATIVE}$ (der Student $\rightarrow$ dem Studenten)
    - $+\text{FUTURE}$ (ser $\rightarrow$ será)
Morphology

- **Clitics**
  - Some words attach to other words.
  - But, orthographic conventions differ:
    - the boy
    - *alwalad* (the boy)

  - She hit him.
  - *darabat thu*. (She hit him.)
A field guide to morphology

Analytic/Isolating

Chinese  English  Spanish  Italian  French
Czech  Polish  Russian  Welsh  Irish  German
Chinese  English  Spanish  Italian  French
Czech  Polish  Russian  Welsh  Irish  German
Synthetic

Maltese  Arabic  Hebrew  Turkish  Finnish  Hungarian
Basque  Navaho  Inuktitut  Mohawk

April 20, 2007
Analytic languages

- No inflectional (category-preserving) morphology
- Some derivational (esp. compounding) morphology

明天 我 的 朋友 为 我 做 生日 蛋糕
míntīn wǒ de péngyou wéi wǒ zuò shēngrì dàngāo
tomorrow I ‘s friend(s) for I to make birthday cake

“My friends will make me a birthday cake tomorrow.”
Fusional languages

• Fusional
  • Most Indo-European languages.
  • Many functional morphological elements (eg. tense, number, gender) combined into a single morpheme.
  • She sings. +s = singular, present tense, indicative
Agglutinative languages

- Agglutinative
  - Hungarian, Finnish, Turkish
  - Concatenate chains of (mostly functional) morphemes

Uygar-laş-tır-a-ma-dık-lar-ımiş-dan-mı-sınız?

Civilized-VERB-CAUS-ABLE-NEG-NOM-PLU-POS1P-ABL-INT-2PL.AGR

“Are you from the ones we could not civilize?”
Polysynthetic languages

- One word, many morphemes
  
  \textit{aliiku-sersu-i-llammas-sua-a-nerar-ta-ssa-galuar-paal-li}

  “However, they will say that he is a great entertainer.”

- A single word may include several open- and closed-class morphemes

  \textit{aliiku} = entertainment \quad \textit{a} = say
  \textit{sersu} = provide \quad \textit{llamas} = good at
Morphology & MT

So why, as MT researchers, do we care about morphology?

1. Inflectional richness → free word order

2. Data sparseness
Morphology & MT

• So why, as MT researchers, do we care about morphology?

  1. Inflectional richness → free word order

  2. Data sparseness
Prior work

- Goldwater & McClosky (2005)
  - Czech → English
  - Preprocess the corpus to throw away some morphemes:
    - Word truncation (ask F.J. Och)
    - Lemmatize everything
    - Only lemmatize infrequent words
    - Keep inflectional morphemes that “mean something” in English
  - Experimentation necessary to determine best process!
Prior work

- Goldwater & McClosky (2005) results:

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-to-word</td>
<td>.311</td>
<td>.270</td>
</tr>
<tr>
<td>lemmatize all</td>
<td>.355</td>
<td>.299</td>
</tr>
<tr>
<td>except Pro</td>
<td>.350</td>
<td></td>
</tr>
<tr>
<td>except Pro, V, N</td>
<td>.346</td>
<td></td>
</tr>
<tr>
<td>lemmatize $n &lt; 50$</td>
<td>.370</td>
<td>.306</td>
</tr>
<tr>
<td>truncate all</td>
<td>.353</td>
<td>.283</td>
</tr>
</tbody>
</table>

*BLEU scores with 5 reference translations, word-based SMT system.
Prior work

- However, with a phrase-based translation model and more data, things look a bit different:

<table>
<thead>
<tr>
<th>Input</th>
<th>BLEU*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>22.81</td>
</tr>
<tr>
<td>Truncated (l=6)</td>
<td>22.07</td>
</tr>
<tr>
<td>Lemmas</td>
<td>22.14</td>
</tr>
</tbody>
</table>

* 1 reference translation, WMT07 dev-test

$p<.05$
Prior work

- What happened?
  - The morphemes that were thrown away had useful information
  - Must avoid *two* pitfalls

Data Sparseness → Information Loss

A Better Translation
Prior work

- Talbot and Osborne (2006)
  - Learn “redundancies” automatically from a parallel corpus
  - Only collapse distinctions that are meaningless w.r.t. a particular target language

- Experiments
  - Smooth surface translation table with revised probabilities
  - Use “compressed” lexicon just to improve word alignments
Prior work

  - Backoff models for machine translation
  - If you don’t know how to translate a word, perform morphological simplification
  - Experiments on Finnish & German
    - German
      - fusional morphology
      - productive compounding
    - Finnish
      - agglutinative morphology
      - Limited noun-noun compounding

Donaudampfschifffahrtsgesellschaften

- Seen? yes, translate
- no, stem:
  - Donaudampfschifffahrtsgesellschaft

- Seen? yes, translate
- no, split compound into 2 pieces
  - Donau Dampfschifffahrtsgesellschaft
Yang & Kirchhoff (2006)

<table>
<thead>
<tr>
<th>Training data</th>
<th>baseline</th>
<th>backoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k</td>
<td>15.3</td>
<td>16.3</td>
</tr>
<tr>
<td>50k</td>
<td>20.3</td>
<td>20.7</td>
</tr>
<tr>
<td>751k</td>
<td>24.8</td>
<td>25.1</td>
</tr>
</tbody>
</table>

**GERMAN**

<table>
<thead>
<tr>
<th>Training data</th>
<th>baseline</th>
<th>backoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k</td>
<td>12.9</td>
<td>14.0</td>
</tr>
<tr>
<td>50k</td>
<td>15.6</td>
<td>16.4</td>
</tr>
<tr>
<td>751k</td>
<td>22.0</td>
<td>22.3</td>
</tr>
</tbody>
</table>

**FINNISH**

- Potential Problems
  - Everything is done as preprocessing
  - Only back off if $C(f) = 0$
  - No improved word alignment
Prior work: take-away

- Morphological simplification can help.
- Morphological simplification can hurt.
  - Only collapse meaningless distinctions!
  - Use a backoff strategy!
- All approaches presented involve making decisions about the translation forms in advance of decoding.
  - Question: Is this the best strategy?
Spoken Language Translation

- Recognize speech in the source language
  - ASR is not perfect!

- Translate into English
  - Translation is not perfect!

- Can we minimize error compounding?
What SLT research tells us

- Joint models better perform better than translating the 1-best hypothesis
  - Ney (1999), Bertoldi et al. (2005a, 2007), Shen et al. (2006)

- Enumerating all hypotheses is not necessary
  - Confusion networks in phrase-based decoders (Moses), Bertoldi (2005a), Bertoldi et al. (2007)
Idea

Model the backoff problem to make it look like speech translation.
The noisy channel

Decoding:

$$\arg\max_e P(e \mid f) = \arg\max_e P(f \mid e)P(e)$$
A noisier channel

Approximation:

\[ S(f') \approx F \]

Decoding:

\[ \arg \max_{e} \max_{f' \in S(f)} P(e, f' | f) \]
Constructing a translation system

• What is $S(f)$?
  • Set of sentences
    • All morphological “alternatives” to $f$ that the system might know how to translate
  • Cost function from a sentence to some value
    • ~How much information did we throw away?

• Constructing $S(f)$
  • Use existing morphological analyzers
  • Truncation
  • Compound splitting
Example

- Given the observed Spanish sentence: *la mujer vieja, S(f)* might contain:

<table>
<thead>
<tr>
<th>SENTENCE</th>
<th>PENALTY</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>la mujer vieja</em></td>
<td>?</td>
</tr>
<tr>
<td><em>EL mujer vieja</em></td>
<td>?</td>
</tr>
<tr>
<td><em>la mujer VIEJ</em></td>
<td>?</td>
</tr>
<tr>
<td><em>EL mujer VIEJ</em></td>
<td>?</td>
</tr>
</tbody>
</table>
Example

- What to do with the penalty?
  - Posterior probability of the sentence under some model (e.g. ASR/OCR word lattices)
  - Amount of morphological information thrown away
    - Count
    - Quantified under some model (e.g. Talbot & Osborne 2006)
  - Function of $#(f)$ vs. $#(g(f))$ in the training corpus
Representing $S(f)$

- $S(f)$ is a huge list with scores! We’d like a compact representation of a huge list.
- Start simple: inflectional morphology
  - Single stem affected
- Confusion networks
  - Good at representing alternatives at a given position
  - Plus, we know how to decode them!
Czech-English translation

- Czech is a highly inflected fusional language.
- Not much compounding.

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Types</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>1.2M</td>
<td>88037</td>
<td>42341</td>
</tr>
<tr>
<td>cz-lemmas*</td>
<td>“</td>
<td>34227</td>
<td>13129</td>
</tr>
<tr>
<td>cz-truncated</td>
<td>“</td>
<td>37263</td>
<td>13039</td>
</tr>
<tr>
<td>English</td>
<td>1.4M</td>
<td>31221</td>
<td>10508</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.4M</td>
<td>47852</td>
<td>20740</td>
</tr>
<tr>
<td>French</td>
<td>1.2M</td>
<td>38241</td>
<td>15264</td>
</tr>
<tr>
<td>German</td>
<td>1.4M</td>
<td>75885</td>
<td>39222</td>
</tr>
</tbody>
</table>

Confusion networks

- CN representation of $S(f)$
  - Surface and lemma at each position
  - Simple penalty model: surface=0, lemma=1

<table>
<thead>
<tr>
<th>z</th>
<th>amerického</th>
<th>břehu</th>
<th>atlantiku</th>
<th>se</th>
<th>veskerá</th>
<th>taková</th>
<th>odůvodnění</th>
<th>jeví</th>
<th>jako</th>
<th>naprosto</th>
<th>bizarní</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>americký</td>
<td>břeh</td>
<td>atlantik</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- atlantiku
- atlantik
Estimating a translation model

- $S(f)$ contains sentences that are a mixture of lemmas and surface forms
- Need translation model that contains both
Estimating a translation model

- Simple solution:
  - Train independent models in parallel
    - Surface → Surface
    - Lemma → Surface
  - Then merge or have two phrase tables available
  - Decoder to chooses the path/translation it likes best
- Pros: easy to estimate
- Cons: except within limits, mixed phrases do not exist!

- A variety of other model possibilities exist!
Czech-English results

<table>
<thead>
<tr>
<th>Input</th>
<th>BLEU*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface forms only</td>
<td>22.74</td>
</tr>
<tr>
<td>Backoff (~Y&amp;K ‘06)</td>
<td>23.94</td>
</tr>
<tr>
<td>Lemmas only</td>
<td>22.50</td>
</tr>
<tr>
<td>Surface+Lemma (CN)</td>
<td>25.01</td>
</tr>
</tbody>
</table>

- Improvements are significant at $p<.05$; CN > surface at $p<.01$.

- WMT07 training data (2.6M words), trigram LM

* 1 reference translation
Czech-English results

Surface only:

From the **US** side of the Atlantic all such *odůvodnění* appears to be a totally bizarre.

Lemma only:

From the *[US]* side of the Atlantic *with* any such *justification* seem completely bizarre.

Confusion Net (Surface+Lemma):

From the **US** side of the Atlantic all such *justification* appears to be a totally bizarre.
Representing other forms of ambiguity

- CNs are fine for inflection, but what about a language with compound/clitic splitting?

\[
gesamthaushaltsplans
\]
\[
gesamthaushaltsplan
\]
\[
gesamt haus halt plans
\]
\[
gesamt haus halt plan
\]

Different lengths!
Confusion nets: the problem

- Every path must pass through every node

<table>
<thead>
<tr>
<th>gesamthaushaltsplans</th>
<th>ε</th>
<th>ε</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>gesamthaushaltsplan</td>
<td>haus</td>
<td>halt</td>
<td>plans</td>
</tr>
<tr>
<td>gesamt</td>
<td></td>
<td></td>
<td>plan</td>
</tr>
</tbody>
</table>
Word lattices

- Any set of strings can be represented
- Algorithms exist for minimizing their size
Decoding word lattices I: Create a chart from the lattice*

- Number nodes by distance from start-node
- For each edge leaving node $i$ and labeled with word $w$, place word $w$ into column $i$
- Augment cell with *span length* (difference between number of next node and current node)

<table>
<thead>
<tr>
<th>gesamthaushaltsplans</th>
<th>4</th>
<th>haus</th>
<th>1</th>
<th>halt</th>
<th>1</th>
<th>plans</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>gesamthaushaltsplan</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>plan</td>
<td>1</td>
</tr>
<tr>
<td>gesamt</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Based on a CKY parser for lattices by Cheppalier (1999)
Decoding word lattices II

- Create translations options for column spans (rather than word spans)
- Column coverage replaces word coverage
- Search for a hypothesis that covers all columns.

  A word may span more than one column!
Decoding word lattices III

<table>
<thead>
<tr>
<th>gesamthaushaltsplans</th>
<th>4</th>
<th>haus</th>
<th>1</th>
<th>halt</th>
<th>1</th>
<th>plans</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>gesamthaushaltsplan</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gesamt</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

----
p=0.0  fc=-100

-**-
p=-10  fc=-40

-***-
p=-15  fc=-75

---*
p=-15  fc=-75

****
p=-15  fc=-75

strategy for the
general budget

****
p=-15  fc=-75

total budget
Word lattice decoding: Problems

- The standard exponential decay distortion model is very poorly defined for word lattices!
  - Lexicalized reordering models fare better.

- Span limits are also poorly defined.
Efficiency of word lattice decoding

- “Morphology” lattices are compact
  - Many nodes that all paths pass through (quasi-linear networks)
  - ASR word lattices do not necessarily have this property!
- Running time proportional to the length of the longest path
## Efficiency of word lattice decoding

### WMT06 German→English Test-Set Stats

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Length</th>
<th>Paths</th>
<th>Decoding time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surface</strong></td>
<td>(27.8)</td>
<td>27.8</td>
<td>1</td>
<td>43 sec/sent</td>
</tr>
<tr>
<td><strong>Split</strong></td>
<td>(31.4)</td>
<td>31.4</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td><strong>Lattice</strong></td>
<td>40.7</td>
<td>31.4</td>
<td>$1.7\times10^9$</td>
<td>52 sec/sent</td>
</tr>
</tbody>
</table>
German-English

- German
  - Fusional inflection (handful of forms)
  - Considerable productive compounding

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Types</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>14.6M</td>
<td>190k</td>
<td>95k</td>
</tr>
<tr>
<td>-stem</td>
<td>“</td>
<td>155k</td>
<td>82k</td>
</tr>
<tr>
<td>-split*</td>
<td>16.3M</td>
<td>83k</td>
<td>33k</td>
</tr>
<tr>
<td>-stem+split</td>
<td>“</td>
<td>67k</td>
<td>29k</td>
</tr>
<tr>
<td>English</td>
<td>15.3M</td>
<td>65k</td>
<td>24k</td>
</tr>
</tbody>
</table>

German-English

- What to do about the penalty function when you can split compounds and stem?

  Er gab uns Übungsblätter (surface)
  Er gab uns Übungsblatt (stem)
  Er gab uns Übung Blätter (split)
  Er gab uns Übung Blatt (stem+split)

- Ideally, two features (weighted or binary): one for splitting and the other for stemming
Results for Word Lattices

- Europarl German → English
  (WMT06 Shared Task, same as Y&K)

<table>
<thead>
<tr>
<th></th>
<th>BLEU*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface-only</td>
<td>25.55</td>
</tr>
<tr>
<td>Lattice (surface-only training)</td>
<td>25.70</td>
</tr>
<tr>
<td>Lattice (combined models)</td>
<td>25.69</td>
</tr>
</tbody>
</table>

* 1 reference translation
Arabic-English

- Arabic segmentation / tokenization / normalization is commonly reported to help (but this is not uncontroversial)
  
  - alra’iis → al ra’iis
  - sayusaafaru → sawfa yusaafaru

- Does segmentation help? Does it lose some important information?
  - Use word lattices to find out!
### Results for Word lattices

- GALE MT03 Arabic → English

<table>
<thead>
<tr>
<th>Input</th>
<th>BLEU*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsegmented</td>
<td>48.12</td>
</tr>
<tr>
<td>Segmented</td>
<td>49.20</td>
</tr>
<tr>
<td>Seg+Noseg (Lattice)</td>
<td>49.70</td>
</tr>
</tbody>
</table>

* 4 reference translations
Conclusion

- Word lattices and CNs have applications aside from speech recognition.
- Preprocessing decisions, such as backoff, can sometimes be better made by the decoder (cf. Czech-English results)
- How much of a problem is morphological sparseness?
Thank You!

Acknowledgements:

Nicola Bertoldi
David Chiang
Marcello Federico
Philipp Koehn
Adam Lopez
Philip Resnik
Daniel Zeman
Richard Zens