Statistical Machine Translation
Part VI – Dealing with Morphology for Translating to German

Alexander Fraser
Institute for Natural Language Processing
Universität Stuttgart

2012.09.18  Seminar: Statistical MT
NSSNLP, University of Kathmandu
Outline

- (Other) work on bitext involving morphologically rich languages at Stuttgart
- Another word on analyzing German compounds
- Morphological generation of German for SMT

Collaborators: Fabienne Braune, Aoife Cahill, Fabienne Cap, Nadir Durrani, Richard Farkas, Anita Ramm, Hassan Sajjad, Helmut Schmid, Hinrich Schuetze, Florian Schwarck, Renjing Wang, Marion Weller
Hindi to Urdu SMT using transliteration

- Hindi and Urdu are very strongly related languages but written in different scripts
- In a small study we determined that over 70% of the tokens in Hindi can be transliterated directly into Urdu
  - The rest must be (semantically) translated
- We designed a new joint model integrating (semantic) translation with transliteration to solve this problem
German subject-object ambiguity

- Example:
  - German: “Die Maus jagt die Katze”
  - Gloss: The mouse chases the cat
  - **SVO** meaning: the mouse is the one chasing the cat
  - **OVS** meaning: the cat is the one chasing the mouse

- When does this happen?
  - Neither subject nor object are marked with unambiguous case marker
  - In the example, both nouns are feminine, article “die” could be nominative or accusative case
  - Quite frequent: nouns, proper nouns, pronouns possible

- We use a German dependency parser that detects this ambiguity and a projected English parse to resolve it
  - This allows us to create a disambiguated corpus with high precision
General bitext parsing

- We generalized the previous idea to a bitext parsing framework.
- We use rich measures of syntactic divergence to estimate how surprised we are to see a triple \((\text{English}\_\text{tree}, \text{German}\_\text{tree}, \text{alignment})\).
  - These are combined together in a log-linear model that can be used to rerank 100-best lists from a baseline syntactic parser.
- New experiments on English to German and German to English both show gains, particularly strong for English to German.
Improved compound analysis for SMT

- Compounds are an important problem for German to English translation and vice versa
- The standard approach to solving this is from Koehn and Knight 2003
- Use a simple linguistic search based on limited linguistic knowledge and the frequencies of words which could form the compound
- We use a high recall rule-based analyzer of German morphology combined with word frequencies to improve beyond this
- Large improvements in METEOR/BLEU beyond Koehn
Example splitting: Ministerpräsident (prime ministre)

Splitting that maximises the score:
Min|ist|Präsident ("Min|is|president")
Example splitting: Ministerpräsident (prime ministre)

Splitting that maximises the score:
Min|ist|Präsident (“Min|is|president”)
Outline

- Work on bitext involving morphologically rich languages at Stuttgart (transliteration, bitext parsing)
- Morphology for German compounds
- **Morphological generation of German for SMT**
  - Introduction
  - Basic two-step translation
    - Translate from English to German stems
    - Inflect German stems
  - Surface forms vs. morphological generation
  - Dealing with agglutination
Early work on morphologically rich languages was the shared task of Romanian/English word alignment in 2005.

I had the best constrained system in the 2005 shared task on word alignment:

- I truncated all English and Romanian words to the first 4 characters and then ran GIZA++ and heuristic symmetrization.
- This was very effective – almost as good as best unconstrained system which used all sorts of linguistic information (Tufis et al).
Early work on morphologically rich languages was the shared task of Romanian/English word alignment in 2005.

I had the best constrained system in the 2005 shared task on word alignment.

- I truncated all English and Romanian words to the first 4 characters and then ran GIZA++ and heuristic symmetrization.
- This was very effective – almost as good as best unconstrained system which used all sorts of linguistic information (Tufis et al).

This alienated people interested in both modeling and (non-simplistic) linguistic features.

- I redeemed myself with the (alignment) modeling folks later.
- Hopefully this talk makes linguistic features people happy.
Morphological Generation of German - Introduction

- For many translation directions SMT systems are competitive with previous generation systems
  - German to English is such a pair
    - The shared task of ACL 2009 workshop on MT shows this
    - Carefully controlled constrained systems are equal in performance to the best rule-based systems
    - Google Translate may well be even better, but we don’t know
      - Data not controlled (language model most likely contains data too similar to test data)
  - English to German is not such a pair
    - Rule-based systems produce fluent output that is currently superior to SMT output
Stuttgart WMT 2009 systems

- German to English system
  - Aggressive morphological reduction (compound splitting & stemming)
  - Deterministic clause reordering using BitPar syntactic parser
  - Worked well (best constraint system)

- English to German system
  - Two independent translation steps
    - Translation from English to morphologically simplified German
    - Translation from morphologically simplified German to fully inflected German
  - Did not work well (worst constraint system)
    - Better modeling is necessary...
Morphological reduction of German

- Morphological reduction driven by sub-word frequencies
  - Simultaneously reduce compounds and stem
  - Compound reduction used Koehn and Knight 2003
  - But it was different: stemming is aggressive; ambiguous suffixes were stripped (motivated by sparsity of news data)

- English to German system tried to invert this process
  - Generate inflected forms (using a second SMT system that translated from reduced representation to normal words using only lemmas and split compounds)
  - This is too hard!
Morphological generation for German

- Goal: fluent output for translation to German
- Problem: German is morphologically rich and English is morphologically poor
  - Many features of German can not be determined easily from English
  - We will focus on 4 features which are primarily aimed at improving NP and PP translation
  - These features are: **Gender, Case, Number, Definiteness**

He sees → Er sieht
a peruvian → einen peruanischen
man → Mann
He sees → Er sieht
a peruvian → einen peruanischen
woman → Frau
Inflection Features

- Gender, Case, Number, Definiteness
  - Diverse group of features
  - Number of the noun and Definiteness of the article are (often easily?) determined given the English source and the word alignment
  - Gender of the noun is innate
    - Often a grammatical gender (for example: inanimate objects in German have genders that are often hard to determine, unlike many Spanish or French nouns)
  - Case is difficult, for instance, often a function of the slot in the subcategorization frame of the verb
  - There is agreement in all of these features in a particular NP
    - For instance the gender of an article is determined by the head noun
    - Definiteness of adjectives is determined by choice of indefinite or definite article
    - Etc...
Overview of translation process

- In terms of translation, we can have a large number of surface forms
- English “blue” -> blau, blaue, blauer, blaues, blauen
- We will try to predict which form is correct
- Our system will be able to generate forms which were not seen in the training data
- We will follow a two-step process:
  1. Translate to “blau” (stem)
  2. Predict features (e.g., Nominative, Feminine, Singular, Definite) to generate the correct form “blaue”
  3. I will compare this with directly predicting “blaue” (e.g. the work presented by Ondrej)
Pros/Cons of 2 step process

- **Pros**
  - Morphological reduction for translation step – better learning from limited parallel data
  - Some inflection is not really a function of English – e.g., grammatical gender. Can predict this using only the German sequence of stems
  - Inflectional features can be treated as something like a (POS) tagging problem
    - Can build tagging system on clean German text with relevant features removed
    - Test it by trying to predict original forms
  - We are solving two easier sub-problems!
Pros/Cons of 2 step process

- **Cons**
  - Conditionality of generation – translate to stems, then predict inflection based on stems
    - No influence of final word forms on stems
    - This is particularly a problem for Case (Case would be difficult anyway, but lexical clues would help)
  - Using features like Case, Definiteness, etc., could be viewed as solving a more difficult problem then necessary
    - We may be modeling definiteness even when it doesn’t matter to generation, etc
Syntactic processing

- Preprocess data:
  - Parse all German data (German side of parallel corpus and German language modeling data) with BitPar, extract morphological features
  - Lookup surface forms in SMOR
  - Resolve conflicts between parse and SMOR
  - Output “stems” (+markup, this will be discussed later) for stem-based translation system

- We also slightly regularize the morphology of English to be more similar to German
  - We use an English morphological analyzer and a parser to try to disambiguate singular/plural/possessive/us (as in Let’s)
  - a/an is mapped to indef_determiner
  - We would do more here if translating, say, Arabic to German
Translating stems

- Build standard phrase-based SMT system
  - Word alignment, phrase-based model estimation, LM estimation
- Run minimum error rate training (MERT)
  - Currently optimizing BLEU on stems (not inflected)
Stem markup

- We are going to use a simple model at first for „propagating“ inflection.
- So we will make some of the difficult decisions in the stem translation step.
- The best German stem markup so far:
  - Nouns are marked with gender and number.
  - Pronouns are nominal or not_nominal.
  - Prepositions are annotated with the case they mark.
  - Articles are only marked definite or indefinite.
  - Verbs are fully inflected.
  - Other words (e.g., adjectives) are lemmatized.
Comparing different stem+markup representations

- BLEU score from MERT on dev (this is abusing BLEU!!)
  - Baseline: 13.49
  - WMT 2009: 15.80
    - Based on Koehn and Knight. Aggressive stemming, reduced compounds. No markup.
  - Initial: 15.54
    - Based on SMOR. Nouns marked with gender and number; coarse POS tag in factored model. No compound handling (will discuss a special case later)
  - “version 1a”: 15.21
    - Same, plus prepositions are marked with case (very useful for ambiguous prepositions)
Review – first step

- Translate to stems
  - But need markup to not lose information
  - This is true of pivot translation as well
- In the rest of the talk I will talk about how to predict the inflection given the stemmed markup
  - But first let me talk about previous work...
Previous work

The two-step translation approach was first tried by Kristina Toutanova’s group at MSR (ACL 2008, other papers)

- They viewed generating an Arabic token as a two-step problem
  - Translate to a sequence of „stems“ (meaning the lemma in Buckwalter)
  - Predict the surface form of each stem (meaning a space-separated token)
- We are interested in two weaknesses of this work
  1. They try to directly predict surface forms, by looking at the features of the surface form
     - I will show some evidence that directly predicting surface forms might not be a good idea and argue for a formal morphological generation step
     - This argument applies to Ondrej’s work as well (I think)
  2. Also, Arabic is agglutinative! Thinking of the token meaning and-his-brother as an inflection of brother is problematic (think about what the English correspondence looks like!)
# Inflection Prediction

<table>
<thead>
<tr>
<th>output decoder</th>
<th>input prediction</th>
<th>output prediction</th>
<th>inflected forms</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>haben&lt; VAPIN &gt;</td>
<td>haben-V</td>
<td>haben-V</td>
<td>haben</td>
<td>have</td>
</tr>
<tr>
<td>Zugang&lt; +NN&gt;&lt;Masc&gt;&lt;Sg&gt;</td>
<td>NN-Sg-Masc</td>
<td>NN-Masc.Acc.Sg.notdef</td>
<td>Zugang</td>
<td>access</td>
</tr>
<tr>
<td>zu&lt;APPR&gt;&lt;Dat&gt;</td>
<td>APRR-zu-Dat</td>
<td>APRR</td>
<td>zu</td>
<td>to</td>
</tr>
<tr>
<td>die&lt; +ART&gt;&lt;Def&gt;</td>
<td>ART-def</td>
<td>ART-Neut.Dat.Sg.def</td>
<td>dem</td>
<td>the</td>
</tr>
<tr>
<td>betreffend&lt; +ADJ&gt;&lt;Pos&gt;</td>
<td>ADJA</td>
<td>ADJA-Neut.Dat.Sg.def</td>
<td>betreffenden</td>
<td>respective</td>
</tr>
<tr>
<td>Land&lt; +NN&gt;&lt;Neut&gt;&lt;Sg&gt;</td>
<td>NN-Sg-Neut</td>
<td>NN-Neut.Dat.Sg.def</td>
<td>Land</td>
<td>country</td>
</tr>
</tbody>
</table>
Solving the prediction problem

- We can use a simple joint sequence model for this (4-gram, smoothed with Kneser-Ney)
- This models $P(stems, coarse-POS, inflection)$
  - Stems and coarse-POS are always observed
  - As you saw in the example, some inflection is also observed in the markup
  - Predict 4 features (jointly)
  - We get over 90% of word forms right when doing monolingual prediction (on clean text)
  - This works quite well for Gender, Number and Definiteness
  - Does not always work well for Case
  - Helps SMT quality (results later)
Surface forms vs morphological generation

- The direct prediction of surface forms is limited to those forms observed in the training data, which is a significant limitation.

- However, it is reasonable to expect that the use of features (and morphological generation) could also be problematic:
  - Requires the use of morphologically-aware syntactic parsers to annotate the training data with such features.
  - Additionally depends on the coverage of morphological analysis and generation.

- Our research shows that prediction of grammatical features followed by morphological generation (given the coverage of SMOR and the disambiguation of BitPar) is more effective.

- This is a striking result, because in particular we can expect further gains as syntactic parsing accuracy increases!
1 LM to 4 CRFs

- In predicting the inflection we would like to use arbitrary features.
- One way to allow the use of this is to switch from our simple HMM/LM-like model to a linear-chain CRF.
- However, CRFs are not tractable to train using the cross-product of grammatical feature values (e.g., Singular.Nominal.Plural.Definite).
  - Using Wapiti (ACL 2010) – Chris says we should be using CDEC...
- Fortunately, we can show that, given the markup, we can predict the 4 grammatical features independently!
- Then we can scale to training four independent CRFs.
Linear-chain CRF features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>lemma_{w_{i-5} \ldots w_{i+5}} \cdot tag_{w_{i-7} \ldots w_{i+7}}</td>
</tr>
<tr>
<td>Case</td>
<td>case_{w_{i-5} \ldots w_{i+5}}</td>
</tr>
<tr>
<td>Gender</td>
<td>gender_{w_{i-5} \ldots w_{i+5}}</td>
</tr>
<tr>
<td>Number</td>
<td>number_{w_{i-5} \ldots w_{i+5}}</td>
</tr>
<tr>
<td>Def.</td>
<td>def_{w_{i-5} \ldots w_{i+5}}</td>
</tr>
</tbody>
</table>

- We use up to 6 grams for all features except tag (where we use 8 grams)
- The only transition feature used is the label bigram
- We use L1 regularization to obtain a sparse model
English features

- SMT is basically a target language generation problem
- It seems to be most important to model fluency in German (particularly given the markup on the stems)
- However, we can get additional gain from prediction from the English, it is easy to add machine learning features to the CRF framework
- As a first stab at features for predicting a grammatical feature on a German word, we use:
  - POS tag of aligned English word
  - Label of highest NP in chain of NPs containing the aligned word
  - Label of the parent of that NP
- Labels: Charniak/Johnson parser then the Seeker/Kuhn function labeler
Dealing with agglutination

- As I mentioned previously, one problem with Toutanova’s work is treating agglutination as if it is inflection.
- It is intuitive to instead segment to deal with agglutination.
- We are currently doing this for a common portmanteau in German:
  - Preposition + Article
  - E.g., „zum“ -> this is the preposition „zu“ and the definite article „dem“
- This means we have to work with a segmented representation (e.g., zu+Dative, definite_article in the stemmed markup) for training and inflection prediction.
  - Then synthesize: creation of portmanteaus dependis on the inflection decision.
- Recently, we got this to work for German compounds as well.
  - We translate to compound head words and compound non-head words, then subsequently combine them. Finally we inflect them.
Evaluation

- WMT 2009 English to German news task
- All parallel training data (about 1.5 M parallel sentences, mostly Europarl)
- Standard Dev and Test sets
- Two limitations of the experiments here:
  - We were not able to parse the monolingual data, so we are not using it (except in one experiment...)
  - The inflection prediction system that predicts grammatical features does not currently have access to an inflected word form LM
- We have recently overcome these, see our EACL 2012 paper
<table>
<thead>
<tr>
<th>System</th>
<th>BLEU (end-to-end, case sensitive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>12.62</td>
</tr>
<tr>
<td>1 LM predicting surface forms, no portmanteau handling</td>
<td>12.31</td>
</tr>
<tr>
<td>1 LM predicting surface forms (11 M sentences inflection prediction training), no portmanteau handling</td>
<td>12.72</td>
</tr>
<tr>
<td>1 LM predicting surface forms</td>
<td>12.80</td>
</tr>
<tr>
<td>1 LM predicting grammatical features</td>
<td>13.29</td>
</tr>
<tr>
<td>4 LMs, each predicting one grammatical feature</td>
<td>13.19</td>
</tr>
<tr>
<td>4 CRFs, German features only</td>
<td><strong>13.39</strong></td>
</tr>
<tr>
<td>4 CRFs, German and English features</td>
<td><strong>13.58</strong></td>
</tr>
</tbody>
</table>
Newest developments

- We now have a rule-based preprocessing setup for English to German translation
  - See our EACL 2012 paper
  - This does reordering of English clauses by analyzing what the translated German clause type will be
- We are currently working on combining inflection, compounding, verbal reordering and verbal morphology prediction
Summary of work on translating to German

- Two-step translation (with good stem markup) is effective
- Predicting morphological features and generating is superior to surface form prediction
  - This depends on quality of SMOR (morph analysis/generation) and BitPar (used for morphological disambiguation here)
  - Performance will continue to improve as syntactic parsing improves
- Linear-chain CRFs good for predicting grammatical features
  - However, tractability is a problem
  - You can get (small gains) with very simple English features
  - More feature engineering work is in progress
Conclusion

- Lecture 1 covered background, parallel corpora, sentence alignment, evaluation and introduced modeling
- Lecture 2 was on word alignment using both exact and approximate EM
- Lecture 3 was on phrase-based modeling and decoding
- Lecture 4 was on log-linear models and MERT
- Lecture 5 briefly touched on new research areas in word alignment, morphology and syntax
- Lecture 6 presented work on translation to German which is relevant to morphologically rich languages in general
Thank you!
General bitext parsing

- Many advances in syntactic parsing come from better modeling
  - But the overall bottleneck is the size of the treebank
- Our research asks a different question:
  - Where can we (cheaply) obtain additional information, which helps to supplement the treebank?
- A new information source for resolving ambiguity is a translation
  - The human translator understands the sentence and disambiguates for us!
Parse reranking of bitext

- Goal: use English parsing to improve German parsing
- Parse German sentence, obtain list of 100 best parse candidates
- Parse English sentence, obtain single best parse
- Determine the correspondence of German to English words using a word alignment
- Calculate **syntactic divergence** of each German parse candidate and the projection of the English parse
- Choose probable German parse candidate with low **syntactic divergence**
Rich bitext projection features

- We initially worked on this problem in the German to English direction
  - Defined 36 features by looking at common English parsing errors
  - Later we added three additional features for the English to German direction
- No monolingual features, except baseline parser probability
- General features
  - Is there a probable label correspondence between German and the hypothesized English parse?
  - How expected is the size of each constituent in the hypothesized parse given the translation?
- Specific features
  - Are coordinations realized identically?
  - Is the NP structure the same?
- Mix of probabilistic and heuristic features
- This approach is effective, results using English to rerank German are strong
New bitext parsing results (not in EACL 2009 paper)

- Reranking German parses
  - This is an easier task than reranking English parses
  - The parser we are trying to improve is weaker (German is hard to parse, Europarl and SMULTRON are out of domain)
  - 1.64% F1 improvement currently, we think this can be further improved

- In the other direction (reranking English parses using a single German parse), we improve by 0.3% F1 on the Brown reranking parser
  - Harder task - German parser is out of domain for translation of the Penn treebank, German is hard to parse. English parser is in domain
Compound Processing: SMOR
Schmid et al. 2004

- finite-state based morphological analyser for German
- covering inflection, derivation and compounding
- good coverage: huge lexicon (over 16,000 noun stems)

Example analysis: Durchschnittsauto ("average car")

```
Durchschnittsauto [NN]
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Durchschnitt</td>
</tr>
<tr>
<td>[NN]</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Auto</td>
</tr>
<tr>
<td>[NN]</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>durchschneiden</td>
</tr>
<tr>
<td>[V]</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>durch</td>
</tr>
<tr>
<td>[VPart]</td>
</tr>
<tr>
<td>schneiden</td>
</tr>
<tr>
<td>[V]</td>
</tr>
</tbody>
</table>
```

SMOR with word frequency results

- Improvement of 1.04 BLEU/2.12 Meteor over no processing
- Statistically significantly better in BLEU than no processing
- Statistically significantly better in Meteor than no processing, and also than Koehn and Knight
- This is an important result as SMOR will be used (together with the BitPar parser) for morphological generation of German