Morphology and Translation

April 17, 2012

Today's goals

- Have a basic understanding of morphology in languages, along with some of the complexities it introduces for MT
- Look at a few approaches to deal with the problem
  - Stemming
  - Splitting
  - Decoding
    - Leveraging ambiguity: Factored representations
    - Preserving ambiguity: Translation from lattices

Motivation

- To this point, we have treated words as atomic white-space delimited units, with no relationships among them
- This hides a lot of information, since words are related

house $\Leftrightarrow$ *Haus*

houses $\Leftrightarrow$ *Hause*

- ...which information is hidden from the computer

Example

Das	*ist*	*ein*	*kleines*	*Haus*	*.*
174	*19*	*182*	*40626*	*991*	*50*
192	*4*	*19*	*27*	*200*	*49*
That	*is*	*a*	*small*	*house*	*.*

Before we go on

- *Start today by discussing with your neighbor some ways in which this view of words loses information*

Morphology

- *Morphology: the study of the forms of words*
  - Inflectional – words change to reflect grammatical roles
    - e.g., *groß, große, großem, großen, großer, großes*
  - Derivational – shared semantics, often across PsOS
    - e.g., *employ (V), employee, employer (N), employable (JJ)*
- Lemma – the basic, canonical form of the word
- Stem – the shared prefix across inflectional variants
  - e.g., *corr- (Spanish)*

Related problem: tokenization

- Morphology is not the only means by which data are unnecessarily fragmented
- Tokenization is largely a task of splitting off punctuation
  - e.g., house, becomes house␣,
  - “No,” he said. becomes “ No , ” he said .
- A related step, normalization, removes case distinctions, standardizes character sets (e.g., quotations, numerals)
- These are largely deterministic processes that are also important for aggregating statistics, but they are largely artifacts of *written* language

Simple morphology: English

- Words are inflected for
  - case (objective, accusative, genitive)
    I, me, my/mine, 's
  - tense (past, present, or future)
    -ed, -ing, will
  - person (1st, 2nd, 3rd)
    I, you, he/she/they
  - number (singular vs. plural)
    -s

Complex morphology: German

- Inflections of the English definite determiner *the*: <u>*the*</u>
- Inflections of the German *male* definite determiner *der*

Case	Singular			Plural		
	male	fem.	n.	male	fem.	n.
nominative (subject)	der	die	das	die	die	die
genitive (possessive)	des	der	des	der	der	der
dative (indirect object)	dem	der	dem	den	den	den
accusative (direct object)	den	die	das	die	die	die

Figure 2.6 Morphology of the definite determiner in German (in English always *the*). It varies depending on count, case, and gender. Each word form is highly ambiguous: *der* is male singular nominative, but also female singular genitive/dative, as well as plural genitive for any gender.

Really complex morphology: Arabic

- Concept defined by three consonants
- Example inflectional morphology:
  - concept: ktb (*to write*)
kataba	he wrote	*CaCaCa*
katabna	we wrote	*CaCaCna*
katabuu	they wrote	*CaCaCuu*
yaktubu	he writes	*yaCCuCu*
naktubu	we write	*naCCuCu*
yaktabuuna	they write	*yaCCaCuuna*
sayaktubu	he will write	*sayaCCuCu*
sanaktubu	we will write	*sanaCCuCu*
sayaktabuuna	they will write	*sayaCCaCuuna*

http://www.eskimo.com/~ram/arabic_morphology.html

Problems caused by morphology

- **In general**
  - *Data sparsity*: alignments to words in the other language are needlessly divided, fracturing statistics

Morphology creates sparsity

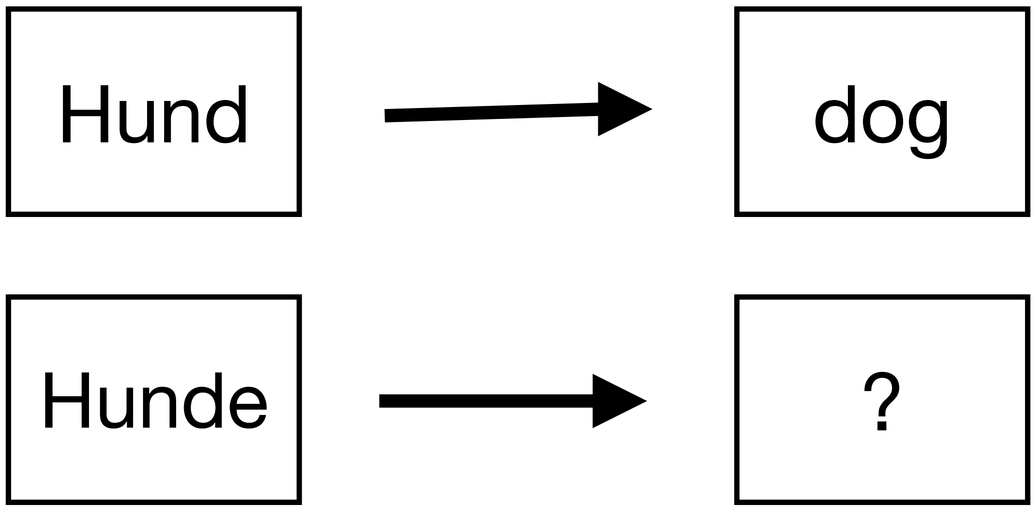
- Common relationships are hidden
  houses = plural(house)
  was = past-tense(is)
  children = plural(child)
- Data is fragmented

	English	German	Finnish
Vocabulary size	65,888	195,290	358,344
Unknown word rate	0.22%	0.78%	1.82%

Figure 10.4 Vocabulary size and effect on the unknown word rate: Numbers reported for 15 million words of the Europarl corpus for vocabulary collection and unknown word rate on additional 2000 sentences (data from the 2005 ACL workshop shared task).

Problems caused by morphology

- In general
  - *Data sparsity*: alignments to words in the other language are needlessly divided, fracturing statistics
- **Source side**
  - *Unseen inflections*: complex inflectional morphology may result in particular versions of a word not being seen



Problems caused by morphology

- In general
  - *Data sparsity*: alignments to words in the other language are needlessly divided, fracturing statistics
- Source side
  - *Unseen inflections*: complex inflectional morphology may result in particular versions of a word not being seen
- **Target side**
  - The right form must be selected, but
    - Richer morphology trades off with word order
    - Morphology can encode long-distance dependencies

Target-side problems

- Inflection varies by case

  I gave her the spoon for her birthday
  Ich gab ihr den Löffel zum Geburtstag

  The spoon was old and rusty
  Der Löffel war alt und rostig

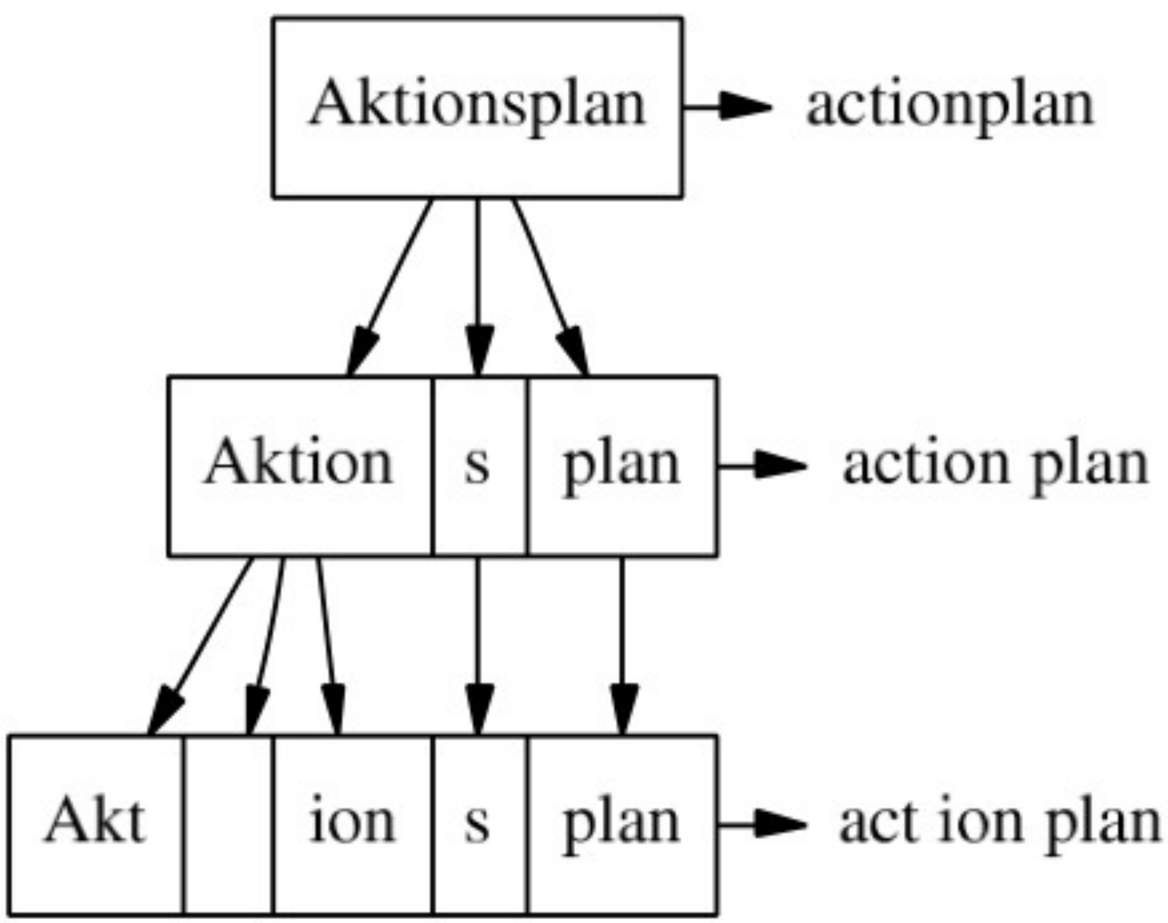
- Inflection frees up word order (in theory, anyway)

Addressing morphology

- There are a number of techniques used to address morphology
  - Splitting
  - Truncation and lemmatization
- And a number of techniques used to incorporate ambiguity and leverage diverse sources of information
  - Decoding from confusion networks
  - Factored translation

Splitting

- An obvious approach is to split up tokens, either manually or automatically
- *Empirical Methods for Compound Splitting (Koehn & Knight, 2003)*



Compound splitting

- German is known for long noun compounds
  - *Großeltern (grandparents)*
  - *Waschmaschine (washing machine)*
  - *Museenverwaltung (museum management)*
- Sometimes this is fine, but sometimes this complicates learning word translations

German-English compound splitting

- *Aktionsplan → action plan, plan of action*
- Technique 1: break word into parts that occur elsewhere
  - aktionsplan (852) = 852
    aktion (960) + plan (710) = 825.6
    aktions (5) + plan (710) = 59.6
    akt (224) + ion (1) + plan (710) = 54.2
- Problem:
  - frei (885) + tag (1864)
    freitag (556)

German-English compound splitting

- Technique 2: make sure parts have translations on the English side
  - since *Frei (free)* and *Tag (day)* are unlikely to exist in the translation of the sentence, *Freitag (Friday)* would not be split
- Problem: ambiguity (the word translations might not always appear)
  - *Grundrechte (basic rights)*
    Grund (reason/foundation) + *rechte (rights)*

German-English compound splitting

- Technique 3: create a separate translation table from the Method 1 technique, use that as a second-level check
- Further issue: common words result in splits
  - *folgenden (following)*
    folgen (consequences) + den (the)
  - solution: POS tag German, limit splitting to certain classes

German-English results

- BLEU score: 30.5 (raw), 34.4 (best splitting)
- Lessons
  - Heuristic splitting is messy: a cascade of exceptions
  - These approaches are also largely specific to German (assuming a particular kind of morphology, and requiring a tagger, for example)

Truncation

- If you don't have a morphological analyzer, a poor man's approximation is to simply truncate the word
  - *What are some limitations of this approach?*
- Goldwater & McClosky (2005) applied this to Czech-English

Words:	`Pro někoho by její provedení mělo smysl .`
Lemmas:	`pro někdo být jeho provedení mít smysl .`
Lemmas+Pseudowords:	`pro někdo být PER_3 jeho provedení mít PER_X smysl .`
Modified Lemmas:	`pro někdo být+PER_3 jeho provedení mít+PER_X smysl .`

Figure 2: Various transformations of the Czech sentence from Figure 1. The pseudowords and modified lemmas encode the verb person feature, with the values 3 (third person) and X ("any" person).

Czech-English

- Truncation isn't as effective as a true lemmatizer, but it's better than nothing

	Dev	Test
word-to-word	.311	.270
lemmatize all	.355	.299
except Pro	.350	
except Pro, V, N	.346	
lemmatize $n < 50$	.370	.306
truncate all	.353	.283

Table 1: BLEU scores for the word-to-word baseline, lemmatization, and word truncation experiments.

Translation from lattices

- In the German-English example, we chose a split for the words prior to learning phrase tables and to translation
- This can be problematic if the segmentation had mistakes
- Idea: preserve the ambiguity of splitting and let the decoder efficiently explore *all* splits

Confusion networks

- A simplified form of lattice
- Czech-English example from Dyer (WMT 2007)

1	2	3	4	5	6	7	8	9	10	11	12
z	amerického americký	břehu břeh	atlantiku atlantik	se s	veskerá	taková takový	odůvodnění	jeví jevit	jako	naprosto	bizarní

Results

- By themselves, lemmatization and truncation were not especially helpful
- A backoff model (in which lower-order models are consulted only when needed) showed some improvement
- The best model made use of a lemmatized confusion network

Input	BLEU	Sample translation
SURFACE	22.74	From the **US side** of the Atlantic all such **odůvodnění** appears to be **a** totally bizarre.
LEMMA	22.50	From the **side** of the Atlantic **with** any such **justification** seem completely bizarre.
TRUNC (l=6)	22.07	From the **bank** of the Atlantic, all such **justification** appears to be totally bizarre.
backoff (SURFACE+LEMMA)	23.94	From the **US bank** of the Atlantic, all such **justification** appears to be totally bizarre.
CN (SURFACE+LEMMA)	**25.01**	From the **US side** of the Atlantic all such **justification** appears to be **a** totally bizarre.
CN (SURFACE+TRUNC)	23.57	From the **US** Atlantic any such **justification** appears to be **a** totally bizarre.

Factored Translation

- Standard phrase-based model: translate sequences of whitespace-delimited tokens
- An alternative is factored translation (Koehn & Hoang, 2007), which simultaneously considers multiples sources of evidence

Factored translation

- Integrates a more complex representation of words directly into the decoder
- Contrast this with some of the other approaches we have considered

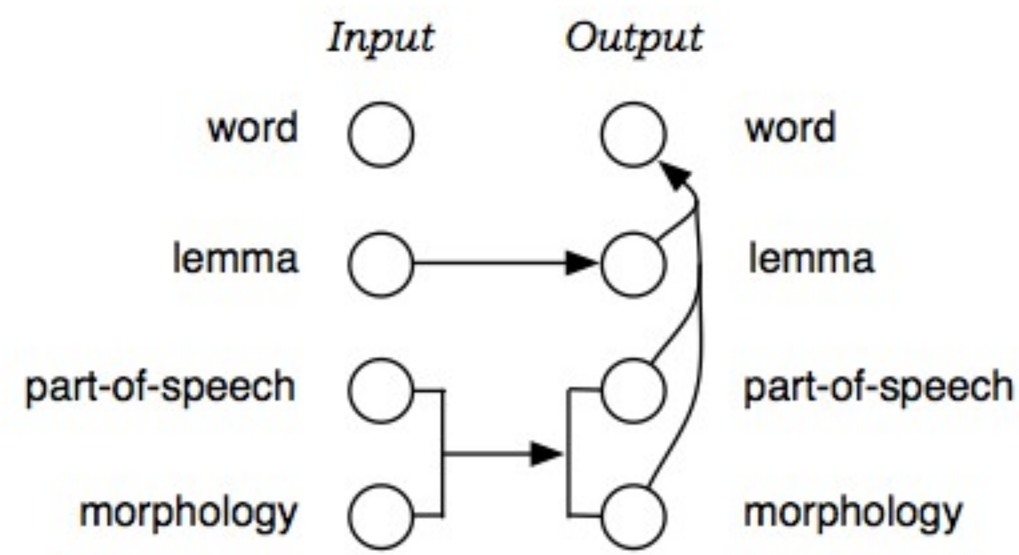


Figure 2: Example factored model: morphological analysis and generation, decomposed into three mapping steps (translation of lemmas, translation of part-of-speech and morphological information, generation of surface forms).

Factored translation

- Steps
  - **Translate** input factors (phrases) into output factors
  - **Generate** surface forms from the output factors (words)
- Example from paper {surface form | lemma | POS | infl}
  - Map lemma {häuser | haus | NN | pl-nom-neut}
    {*?|house|?|?*, *?|home|?|?*,*?|building|?|?*}
  - Map morphology
    {?|house|NN|pl, ?|home|NN|pl,
    ?|building|NN|pl, ?|house|NN|sg}
  - Generate surface
    {houses|house|NN|pl, homes|home|NN|pl,
    buildings|building|NN|pl, house|house|NN|sg}

Results

English–German

Model	BLEU
best published result	18.15%
baseline (surface)	18.04%
surface + POS	18.15%
surface + POS + morph	18.22%

English–Spanish

Model	BLEU
baseline (surface)	23.41%
surface + morph	24.66%
surface + POS + morph	24.25%

English–Czech

Model	BLEU
baseline (surface)	25.82%
surface + all morph	27.04%
surface + case/number/gender	27.45%
surface + CNG/verb/prepositions	27.62%

Summary

- Morphology is a real problem in translation, especially for low-resource languages
- Linguistic approaches are useful (e.g., lemmatization), and even linguistic approximations (e.g., truncating) can do well
- Morphology is far from a solved problem

References

- *Empirical Methods for Compound Splitting* (Koehn & Knight, 2005)
- *The 'noisier channel': translation from morphologically complex languages* (Dyer, WMT 2007)
- *Factored Translation Models* (Koehn & Hoang, 2007)
- *Improving Statistical MT through Morphological Analysis* (Goldwater & McClosky, 2005)