CMPE 409 Machine Translation

Phrase-Based Models

Murat ORHUN
(adapted from the original slides of Prof. Philipp Koehn)

Istanbul Bilgi University

May 18, 2022

Recall:Word based translation

- One to one translation
- One to many translation
- Too many cases to consider

Phrase-Based Models

- Best performing statistical machine translation systems are based on phrase-based models
- Models that translate small word sequences at a time
- Explain basic principles of phrase-based models and how they are trained
- Explain translation model and the reordering model (Introduction).

Motivation

- Word-Based Models translate *words* as atomic units
- Phrase-Based Models translate *phrases* as atomic units

Word-based translation

- One to one translation
- One to many translation
- Too many cases to consider

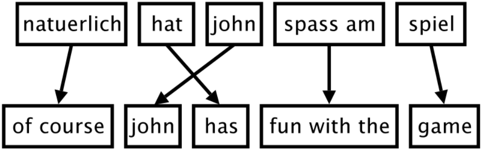
Phrase-based translation

- Many-to-many translation can handle non-compositional phrases
- Use of local context in translation
- The more data, the longer phrases can be learned

Phrase-based translation

"Standard Model", used by Google Translate and others

Phrase-based Model



- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Phrase Translation Table

- Main knowledge source: table with phrase translations and their probabilities
- Example: phrase translations for natuerlich

Translation	Probability $\phi(\bar{e}\|\bar{f})$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Real Example

- Phrase translations for den Vorschlag learned from the Europarl corpus:

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

  - lexical variation (proposal vs suggestions)
  - morphological variation (proposal vs proposals)
  - included function words (the, a, ...)
  - noise (it)

Linguistic Phrases

- Model is not limited to linguistic phrases (noun phrases, verb phrases, prepositional phrases, ...)
- with the game: fun is a noun phrase and with the game is a prepositional phrase

Linguistic Phrases

- Example non-linguistic phrase pair:

$$\text{spass am} \rightarrow \text{fun with the}$$

- Prior noun often helps with translation of preposition:
  - am is usually translated to on the or at the, but with the is rather unusual.
- Experiments show that limitation to linguistic phrases hurts quality.

Probabilistic Model

- Bayes rule

$$\begin{aligned} \mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{\text{LM}}(\mathbf{e}) \end{aligned}$$

  - translation model $p(\mathbf{f}|\mathbf{e})$
  - language model $p_{\text{LM}}(\mathbf{e})$

- Decomposition of the translation model

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1)$$

  - phrase translation probability ϕ
  - reordering probability d

Distance-Based Reordering

- Reordering is relative to the previous phrase:

$$d(start_i - end_{i-1} - 1)$$

  - $start_i$ is the position of the first word of the foreign phrase that translates to the ith English phrase.
  - end_i is the position of the last word of that foreign phrase.
  - end_{i-1} is the position of the last word of the foreign phrase that translates to the $(i-1)$th English phrase.
  - reordering distance is computed as $start_i - end_{i-1} - 1$

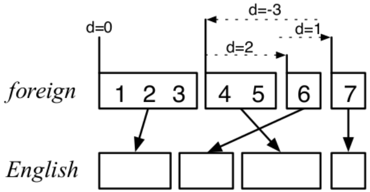
Distance-Based Reordering

- The reordering distance is the number of phrases skipped (forward
backward):

$$d(start_i - end_{i-1} - 1)$$

  - If two phrases are translated in sequence: $start_i = end_{i-1} - 1$ (reordering cost $d(0)$)

Distance-Based Reordering



phrase	translates	movement	distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	-3
4	7	skip over 6	+1

Distance-Based Reordering

- Scoring function: $d(x) = \alpha^{|x|}$ — exponential with distance
- Movements of phrases over large distances are more expensive tha short distances or no movement at all.
- $\alpha \in [0, 1]$

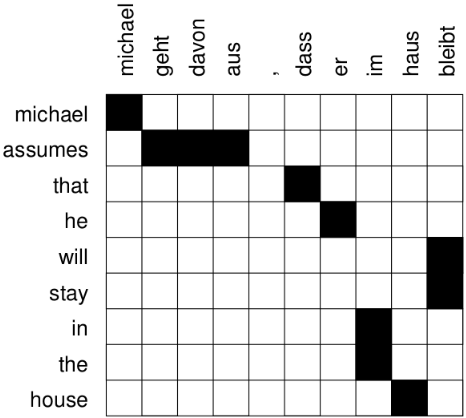
Benefits of phrases over words for translations

- Words may not be the best atomic units, due to one-to-many mappings (and vice-versa).
- Translating words groups helps to resolve ambiguities.
- It is possible to learn longer and longer phrases based on large training corpora.
- We do not need to deal with the complex notions of fertility, insertion and deletions.

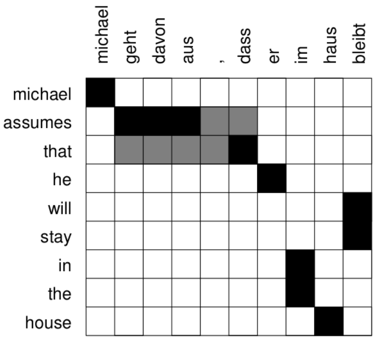
Learning a Phrase Translation Table

- Task: learn the model from a parallel corpus
- Three stages:
  - word alignment: using IBM models or other method
  - extraction of phrase pairs
  - scoring phrase pairs

Word Alignment



Word Alignment

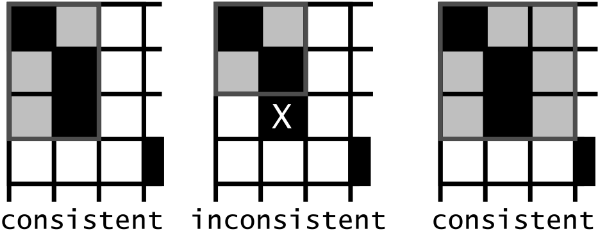


extract phrase pair consistent with word alignment:
assumes that / geht davon aus , dass

Extracting Phrase Pairs

- Phrases can be shorter or longer:
  - Shorter phrases occur frequently and are more often applicable to unseen sentences.
  - Longer phrases capture more local context and can be used to translate large chunks of text at one time.

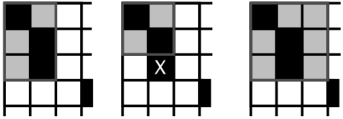
Consistency



ok	violated	ok
	one alignment point outside	unaligned word is fine

All words of the phrase pair have to align to each other.

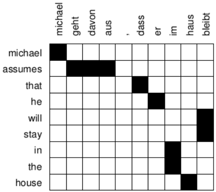
Consistency



Phrase pair $(\bar{e}, \bar{f})$ consistent with an alignment A, if all words $f_1, ..., f_n$ in $\bar{f}$ that have alignment points in A have these with words $e_1, ..., e_n$ in $\bar{e}$ and vice versa:

$$
\begin{aligned}
(\bar{e}, \bar{f}) \text{ consistent with } A \Leftrightarrow \\
& \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\
\text{AND } & \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e} \\
\text{AND } & \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A
\end{aligned}
$$

Phrase Pair Extraction



Smallest phrase pairs:

michael — michael

assumes — geht davon aus / geht davon aus ,

that — dass / , dass

he — er

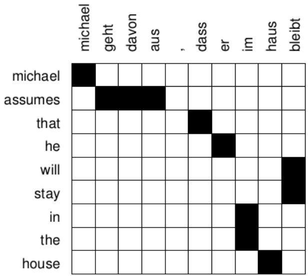
will stay — bleibt

in the — im

house — haus

unaligned words (here: German comma) lead to multiple translations

Larger Phrase Pairs



michael assumes — michael geht davon aus / michael geht davon aus ,
assumes that — geht davon aus , dass ; assumes that he — geht davon aus , dass er
that he — dass er / , dass er ; in the house — im haus
michael assumes that — michael geht davon aus , dass
michael assumes that he — michael geht davon aus , dass er
michael assumes that he will stay in the house — michael geht davon aus , dass er im haus bleibt
assumes that he will stay in the house — geht davon aus , dass er im haus bleibt
that he will stay in the house — dass er im haus bleibt ; dass er im haus bleibt ,
he will stay in the house — er im haus bleibt ; will stay in the house — im haus bleibt

Remarks on Phrase Pair Extraction

- We cannot extract matching German phrases for some English phrases
  - e.g., **im** is mapped to both **in** and **the**
- We cannot extract matching German phrases for some English phrases
  - e.g., **he will stay** cannot be mapped to **er ... bleibt**
- Unaligned words can lead to multiple matches
  - e.g., the comma can be aligned or not together with **dass**

Remarks on Phrase Pair Extraction

- Some statistics:
  - 9 English words, 10 German words: 11 alignment points
  - 36 English phrases, 45 German phrases: 24 pairs extracted
- The number of extracted phrases can be quadratic in the number of words.
- Limiting the length of the phrases is recommended.

Scoring Phrase Translations

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$

Recall: Standard Model

- Phrase-based standard model better than word based model
- Achieves generally better translation
- could be improved with Extension models

Components of standard model

- Described standard model consists of three sub-models:
  - phrase translation model $\phi(\bar{f}|\bar{e})$
  - reordering model d
  - language model $p_{LM}(e)$

$$e_{\text{best}} = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ d(start_i - end_{i-1} - 1) \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1 \ldots e_{i-1})$$

Weighted model

- Some sub-models may be more important than others
- Add weights λ_ϕ, λ_d, λ_{LM}

$$e_{\text{best}} = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)^{\lambda_\phi} \; d(start_i - end_{i-1} - 1)^{\lambda_d} \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1 \ldots e_{i-1})^{\lambda_{LM}}$$

Other models

- Log-Linear Model
- Weighted Model as Log-Linear Model
- Bidirectional translation probabilities
- Lexical weighting
- Word penalty
- Phrase penalty
- Featured Functions

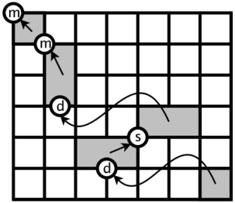
Lexical Reordering

Reordering is one of the hardest problems in machine translation

- Different language pairs need different types of reordering:
  - local: French, Arabian, Chinese to English
  - distant: German, Japanese to English

- Our reordering model generally punishes movement and it is up to the language model (usually based on trigrams) to justify the placement of words in a different order.

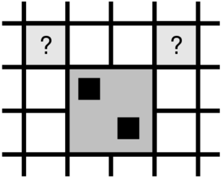
Lexicalized Reordering



- Distance-based reordering model is weak
  $\rightarrow$ learn reordering preference for each phrase pair
- Three orientations types: (m) monotone, (s) swap, (d) discontinuous

$$\text{orientation} \in \{m, s, d\}$$
$$p_o(\text{orientation}|\bar{f}, \bar{e})$$

Learning Lexicalized Reordering



- Collect orientation information during phrase pair extraction
  - if word alignment point to the top left exists → **monotone**
  - if a word alignment point to the top right exists→ **swap**
  - if neither a word alignment point to top left nor to the top right exists → neither monotone nor swap → **discontinuous**

Learning Lexicalized Reordering

- Estimation by relative frequency

$$p_o(\text{orientation}) = \frac{\sum_{\bar{f}} \sum_{\bar{e}} count(\text{orientation}, \bar{e}, \bar{f})}{\sum_o \sum_{\bar{f}} \sum_{\bar{e}} count(o, \bar{e}, \bar{f})}$$

- Smoothing with unlexicalized orientation model $p(\text{orientation})$ to avoid zero probabilities for unseen orientations

$$p_o(\text{orientation}|\bar{f}, \bar{e}) = \frac{\sigma\ p(\text{orientation}) + count(\text{orientation}, \bar{e}, \bar{f})}{\sigma + \sum_o count(o, \bar{e}, \bar{f})}$$

EM Training of the Phrase Model

- What if we do not have the word alignment for the sentences?
- Could we align the phrases directly from the sentence pairs?
- We presented a heuristic set-up to build phrase translation table (word alignment, phrase extraction, phrase scoring)
- Alternative: align phrase pairs directly with EM algorithm

EM Training of the Phrase Model

- EM algorithm:
  - initialization: uniform model, all $\phi(\bar{e}, \bar{f})$ are the same
  - expectation step:
    - estimate likelihood of all possible phrase alignments for all sentence pairs
  - maximization step:
    - collect counts for phrase pairs $(\bar{e}, \bar{f})$, weighted by alignment probability
    - update phrase translation probabilties $p(\bar{e}, \bar{f})$

Drawbacks of EM for Phrase Model

- There are many possibilities of phrases (input and output).
  - We might want to limit the phrases space: minimum occurrences for a phrase or phrase pair.
- Greedy search heuristic: can find high-probability phrase alignments in a reasonable time.
- Results are usually no better than using word alignments as input.
  - The method easily overfits: learns very large phrase pairs, spanning entire sentences.

Summary

- Phrase Model
- Training the model
  - word alignment
  - phrase pair extraction
  - phrase pair scoring
- Log linear model
  - sub-models as feature functions
  - lexical weighting
  - word and phrase count features
- Lexicalized reordering model
- EM training of the phrase model

Further Reading

- Statistical Machine Translation, Philipp Koehn (chapter 5).

https://www.youtube.com/watch?v=AwZSgHikWwk&t=1001s

Recourse

- Jurafsky, D. and J. H. Martin. Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, Upper Saddle River, NJ: Prentice-Hall, 2008.
- Koehn, P. (2009). Statistical Machine Translation. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815829