Machine Translation Systems

Gerald Penn

CS224N / Ling 284

[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky and Chris Manning]

A complete translation system

Decoding for IBM Models

• Of all conceivable English word strings, find the one maximizing \( P(e) \times P(f | e) \)

• Decoding is NP hard
  – (Knight, 1999)

• Several search strategies are available
  – Usually a beam search where we keep multiple stacks for candidates covering the same number of source words
  – Each potential English output is called a hypothesis.

Search for Best Translation

\[ \text{voulez – vous vous taire !} \]

\[ \text{voulez – vous vous taire !} \]

\[ \text{you – you you quiet !} \]

\[ \text{voulez – vous vous taire !} \]

\[ \text{quiet you – you you !} \]
Search for Best Translation

voulez – vous vous taire !

you shut up !

Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

The “Fundamental Equation of Machine Translation” (Brown et al. 1993)

\[
\hat{e} = \arg\max P(e | f)
\]

\[
= \arg\max P(e) \times P(f | e) / P(f)
\]

\[
\ne = \arg\max P(e) \times P(f | e)
\]

Rewards longer hypotheses, since these are ‘unfairly’ punished by P(e)
What StatMT people do in the privacy of their own homes

\[
\arg\max_e P(e)^{1.9} \times P(f \mid e) \times 1.1 \times \text{length}(e) \times KS^{3.7} \ldots
\]

Lots of knowledge sources vote on any given hypothesis.

“Knowledge source” = “feature function” = “score component”.

Feature function simply scores a hypothesis with a real value.

(May be binary, as in “e has a verb”).

Problem: How to set the weights?

(We look at one way later: maxent models.)

---

Flaws of Word-Based MT

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - “real estate”, “note that”, “interested in”
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

---

Phrase-Based Statistical MT

- Foreign input segmented into phrases
  - “phrase” is any sequence of words
- Each phrase is probabilistically translated into English
  - “the conference”\(\rightarrow\) “zur Konferenz”
  - “the meeting”\(\rightarrow\) “zur Konferenz”
- Phrases are probabilistically re-ordered

See J&M or Lopez 2008 for an intro.

This is still pretty much the state-of-the-art!

---

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguating
  - “interest rate”\(\rightarrow\) …
  - “interest in”\(\rightarrow\) …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

---

How to Learn the Phrase Translation Table?

- Main method: “alignment templates” (Och et al., 1999)
- Start with word alignment, build phrases from that.
How to Learn the Phrase Translation Table?

- Main method: “alignment templates” (Och et al. 1999)
- Start with word alignment, build phrases from that.

```
Mary did not slap the green witch
Maria no dió una bofetada a la bruja verde
```

This word-to-word alignment is a by-product of training a translation model like IBM-Model-3. This is the best (or “Viterbi”) alignment.

IBM Models are 1-to-Many

- Run IBM-style aligner both directions, then merge:

```
E->F best alignment
F->E best alignment
```

Collect or abstraction or cleverer algorithm

Phrase Pair Probabilities

- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.
- No EM training
- Just relative frequency:

```
P(f-f-f | e-e-e) = \frac{count(f-f-f, e-e-e)}{count(e-e-e)}
```

Phrase-Based Translation

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.
Phrase-Based Translation

Phrase-Based Translation

Syntax and Semantics in Statistical MT

Why Syntax?
- Need much more grammatical output
- Need accurate control over re-ordering
- Need accurate insertion of function words
- Word translations need to depend on grammatically-related words

Yamada and Knight (2001): The need for phrasal syntax
- He adores listening to music.

Kare ha ongaku wo kiku no ga daisuki desu
**Syntax-based Model**

- E→J Translation (Channel) Model

  ![Parse Tree(E) → Sentence (J)](image)

  1. Preprocess English by a parser
  2. Tokenize by Chasen
  3. Parse by Collins Parser
  4. Training Corpus: J-E 2K sentence pairs
  5. E: Parsed by Collins Parser
  6. EM Training: 20 Iterations

**Experiment**

- Training Corpus: J-E 2K sentence pairs
- J: Tokenized by Chasen (Matsumoto, et al., 1999)
- J: Parsed by Collins Parser (Collins, 1999)
- E: Parsed by Collins Parser (Collins, 1999)
- E: Flatten parse tree
- EM Training: 20 Iterations
- To Capture word-order differences (SVO→SOV)

**Result: Alignments**

- Ave. Score
- # perf. sent.

| Y/K Model | 0.582 | 10 |
| IBM Model | 0.431 | 0 |

- Avg. by 3 humans for 50 sentences
- okay(1.0), not sure(0.5), wrong(0.0)
- precision only
MT Applications

Gerald Penn
CS 224N
2011
[Based on slides by Chris Manning]

MT Applications: 1. Traditional

- Traditional scenario:
  - Documents had to be translated for your company/organization. Document production for organization
  - Generally, the quality/accuracy demands are high
  - High cost
- Though most of it is now done as outsourced piecework
- MT tends to be ineffective: The cost of post-translation error correction is too high
- Main technology in the game: translation memory/translation workbench/terminology management
  - E.g., TRADOS.
- Very slowly, MT technology is starting to be incorporated, but most of the action is in terminology/lexicon management.

MT Applications: 2. Web

- Web applications:
  - Dominant scenario: User-initiated translation
    - Crucial difference: The quality doesn’t have to be great. The user is usually okay with just understanding the gist of what is going on
  - Second scenario
    - Somehow on the web people will accept medium quality results. Accessible information is better than no information
  - MT is saved!!! “It’s the web, stupid.”
    - But is there money in it?
AltaVista
BabelFish
1997:
Free, automatic
translation for the
masses.
Revolutionary.

But, what was the
underlying technology?
SYSTRAN.

MacOS Dashboard?
SYSTRAN
Google until 2006?
SYSTRAN
Machine Translation Summary

- Usable Technologies
  - “Translation memories” to aid translator
  - Low quality screening/web translators
- Technologies
  - Traditional: Systran (Altavista Babelfish, what you got till mid-2006 on Google) is now seen as a limited success
  - Statistical MT over huge training sets is successful (IS/LanguageWeaver, Microsoft, Google)
- Key ideas of the present/future
  - Statistical phrase based models
  - Syntax based models
  - Better language models (e.g., bigger, using grammar)
  - Better decoding models (e.g., by restricting model?)