1. Words, and other words

A brief introduction to machine translation.

Machine Translation covers a wide range of goals:
- Fully automatic high-quality unrestricted MT
- To aid
  - Machine-aided human translation
- Machine-aided human translation
  - Helps in the language learning, but not of the machine-learning kind

The contract between hyped-up promises of success and poor actual performance led to:
- the AMC report (1946)
- Realising that many years of research had failed to meet expectations
  - Unit has no shortage of translation
  - Fully automatic MT hasn’t been worth the effort — quality hasn’t improved much
  - It isn’t clear if it will ever work

Usually we talk about:
- Bilingual
  - For each position in the Russian, choose a translation of the English word which aligns
- Gigaword corpus
  - And a variety of HMM-based decoding methods
- The channel distorts this into what we see or hear, that is, Russian
  - But most MT work today is based on one form or another of noisy channel decoding

For word-word alignment and word/phrase translation:
- Choose a length for the Russian, given the length of the English
  - So, for our trivial Russian example, this would be
  - By the IBM team, working from French into English
  - Using Canadian Hansard for training and testing

4. A very noisy channel

Applying the noisy channel model to translation requires us to stand normal terminology on its head:
- Usually we talk about source and target language
  - For example, when translating English to Japanese
    - Russian is the source
    - English is the target
  - But from the perspective of the noisy channel model:
    - The source is English
    - The channel distorts this into what we see or hear, that is, Russian
    - Which we have to decode
      - To give it to the source
      - Which is the target

5. Priors and likelihood for MT

Remember the basic story (using e for English and r for Russian):

\[
P(r_1^n | e_1^n) \propto \text{prior} \times \text{likelihood}
\]

The prior is just our old friend, some form of language model:
- for English!
- But the channel model needs to be articulated a bit for translation, in several ways:
  - The source and target need not have the same number of words
  - And the mapping part, even in a fairly simple view, has to do two things:
    - But just translate the words
    - But order them as well!

So we need a channel model that takes all of these into account.

6. Translation modeling for MT

[SKIM MT]. Chapter 25 takes you through a step-by-step motivation for the first successful attempt at doing things this way:
- By the IBM team, working free hand into English
  - Using Canadian material for training and testing
  - And a variety of HMM-based decoding methods.

All other approaches start with a formal notion of alignment:
- A (possibly one-to-many) mapping from source word position to position in the Russian
- For our trivial Russian example, this would be
- Because the source and target words exchange positions between English and Russian
  \[ \text{English: Will you save me?} \]
  \[ \text{Russian: Izbylo voskri.} \]

7. Translation modeling, cont’d

Their simplest model then has three conceptual steps:

1. Choose a length for the Russian, given the length of the English
  - Remember, from the corpus perspective we are generating the Russian observations, starting from the English source
  - Basic of the PRF tagging
2. Choose an alignment from the words in the source (English) to the words in the channeled observations from a sequence of POS tags in the target
3. For each position in the Russian, choose a translation of the English word which aligns to it

Following simplifying assumptions of the usual Markov nature, we end up with:

\[
P(r_1^n | e_1^n) = \prod_{j=1}^n P(r_j | r_{j-1}^j, e_1^n)
\]

Where
- \( n \) is the alignment of the \( n \) Russian word
- \( e_1^n \) is the alignment of the \( n \) English word

For the translation model before we do the Bayes rule switch and the \( \text{argmax} \):

8. Contemporary MT

We’ve barely scratched the surface.

But state-of-the-art MT systems today all derive from essentially the starting point:
- Including today’s Parrots
- Which also to think that “я вас любил” should be translated as “I loved you more”

9. Getting started: The role of data

Broadly speaking, we have two models to learn:
- Serve from the dictionary
  - Target language data
  - One-to-one
    - Loss of 1
  - Loss of 1
    - Even though it’s wrong
  - For word alignment and word-to-phrase translation
- Need to get it as well

10. Getting started: segmentation and sentence alignment

Just as with other corpora, we need to pre-process the raw materials:
- Normalise markup
Check for and correct character encoding problems.
Segment and normalise:
- tokens
- morphemes?
- sentences
- paragraphs
- down-casing of beginning of sentences, maybe tag

These will vary in difficulty given the form of the raw data.
- And the language involved.

But for the translation models, with respect to the bilingual data, we need more:
- We need to align the two versions of the paragraph and sentence level
  - Sentence level is not always 1-to-1

11. Sentence alignment details: Gale and Church (1993)

Assumptions:
- We start with true documents
  - In source and target languages
  - Translations of one another
- Sentence order is stable if ever changed
- If paragraphs exist, they are already aligned

Paragraph by paragraph, the algorithm matches source sentences to zero, one or two target sentences:
- Sentence may be divided by translator
- Sentence may be split in half by translator
- In either direction
- We don’t actually always know which was the original

12. Gale and Church, cont’d

Start with some empirical observations:
What does a hand-aligned corpus tell us about sentence alignment?

13. Evaluation-driven development

From 2006–2014, an annual competition was held:
- The ‘Workshop’ on Statistical Machine Translation (WMT)
  - Shared task, many language pairs
- Participants given corpora with which to train their MT systems
- They get a test set to translate and submit
- Judges disagree with each other
- Slow sentences
- Segment and normalise
- Where by
- And the language involved
- And with themselves (from one trial to the next)
- Sentence may be deleted by translator
- Participants write papers on how they built their systems for the conference
- Reference translation(s) (maybe)
- Translations of one another
- Misleading if based on a single translation: there is no one ‘right’ answer
- That gives G&C the basis for a maximum likelihood estimate of
  - Deletion, insertion, substitution
  - With costs coming from the formula above, drawing on
    - That is, you can’t compute mean, variance, etc.

14. Evaluation

How can we evaluate systems?

As with other similar tasks, in one of two ways:

**Intrinsic evaluation**
- For the ability of the result with respect to some (real) use
  - As fast as real production: how much post-editing required?
  - Comprehensiveness
  - As for search or information retrieval: measure quality of that result

**Extrinsic evaluation**
- Measure the quality of the result against some (more-or-less explicit) standard
  - Human-quality assessment
  - Automatic comparison to gold standard

Any intrinsic (human-based) is:
- Bias
- Expensive
- Hard to ensure fairness
- Not stable
- Judges disagree with each other
- Add with hesitation (from one trial to the next)

Any automatic measures is:
- Only as good as the gold standard it uses
- Misleading if based on a single translation, there is no one ‘right’ answer

15. Human evaluation

One or more judges, working in pairs:
- MT system output
- Original
- Reference translation(s) (maybe)

Different dimensions for judgement:
- Error: is the meaning of the source preserved?
- Error: is the result in the target language?
- Typically judged on a numeric scale
  - Which is misleading
  - You can’t trust the results at all
  - That is, you can’t compare mean, variance, etc.

"Valentino má vždycky radši elegantní red sisitu."  
- Score
- Translation
  - Sentence should always elegance rather than fame.
- Sentences

"Valentino has always preferred elegance to notoriety.
- Reference
  - Translation
  - Sentence
  - Sentence
  - Sentence
  - Sentence
  - Sentence

"Valentino has always rather than the elegance of glory.
- Translation
  - Reference
  - Sentence
  - Sentence
  - Sentence

"Valentino has always had the elegance rather than glory.
- Translation
  - Reference
  - Sentence
  - Sentence
  - Sentence
  - Sentence

As mentioned above, agreement can be a problem
16. Automatic evaluation

There was no accepted automatic evaluation measure for MT for a long time.

- Problems with the evaluation-driven funding ideology
- n-gram overlap
- System which makes one and only one guess, and gets it right
- System which guesses the whole lexicon every time

That is, we initialise the translation model with either random, or uniform, estimates of
probability for all possible alignments and all possible translations.

The advent of the BLEU methodology (Bilingual Evaluation Understudy) around 2000 helped a
lot.

- By Roberts and colleagues at IBM T. J. Watson Labs
- It correlates surprisingly well with human judgments
- Although it's nowhere near perfect
- It's good enough for now

17. A digression about headroom

When you need numerical scores to facilitate hillclimbing

- It really matters how far you are from your goal
- If your system is being pretty well already
  - You need a very accurate measure to reliably detect improvement
- But if you're doing pretty badly
  - A rough-and-ready measure will be just fine

So we can ask "How much headroom do we have?"

20. BLEU: Three versions of the formula

As described, the BLEU formula is a product

- Of the brevity penalty
- BP = \( \frac{N_w}{\text{ref}} \), where \( N_w \) is the number of words in the candidate
- And \( \text{ref} \) is the number of words in the reference
- And the geometric mean of the modified n-gram precisions
- \( \sqrt[n]{\text{prec}_n} \), where \( \text{prec}_n \) is the precision for all possible alignments and all possible translations.

The fourth root is usually expressed via the log domain
- \( \text{log} \sum_{n-gram} \text{prec}_n \)

The whole thing is usually then moved into the log domain
- \( \text{log} \frac{\text{ref}}{\text{c}} \)

For simplicity in presentation, as well as the usual practical reasons

\( \text{BLEU} = \text{log} \sum_{n-gram} \text{prec}_n \)

See J&M 2 for details and a worked example

19. A digression about precision and recall

We'll assume that we have at least the beginnings of a bilingual lexicon

Results for a document retrieval request vs. the 'correct' set

- But goes beyond that some way towards checking that they're in the 'right'
- Has 100% recall
- But very low precision

Well, not

And

That's

Giving us something like this

MT has plenty of headroom
- Thus the version above that BLEU is "good enough for now"
What happened with ‘fleur’ is called the pigeon hole principle:

- There’s nowhere else plausible for it to map to.

23. But this is all changing

Over the last two years, there’s been a huge shift in emphasis:

- Away from the explicit noisy channel architecture with multiple components:
  - Hand-tuned, weighted and combined
  - To deep neural nets:
    - Which may have several components:
    - But not as easily distinguish the channel model from the language model.

Google announced a few months ago that Google Translate had made such a shift for some of the most common language pairs.

24. Expectation maximisation, cont’d

As aggregating across all sentence pairs, we can count how much probability attaches to e.g. the bleu/blue pair, compared to all the other... pairs, to get a new ML estimate.

In the simplest IBM model, so-called IBM Model 1, they started with the assumption that all alignments were equally likely.

The overall shape of their EM process was as follows:

Step 1: Expectation

- For every sentence pair:
  - For every possible word alignment:
    - Use the word-word translation probabilities to assign a total probability

Step 2: Maximisation

- Suppose the assigned values to be true:
- Collect probability-weighted counts for all word translation pairs:
- Re-estimate the probabilities for every pair:

Iterate until convergence:
- That is, go back to Step 1 and use the re-estimated word-word probabilities to re-estimate the alignment probabilities.

This approach performed well enough to launch a revolution.

25. Expectation maximisation: simplified example


Assume one-to-one word alignments only.

So we have

Where

- $t(a|f,e)$ is the alignment
- $P(f)$ is a foreign sentence
- $P(e)$ is an English sentence
- $J$ is the length of the foreign sentence

And just two pairs of sentences:

- “black cat”: “chat noir”
- “the cat”: “le chat”

Giving the following vocabularies:

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>black, cat, the</td>
<td>chat, le, noir</td>
</tr>
</tbody>
</table>

26. EM Example, cont’d

We start with uniform word translation probabilities:

| $t(chat|black)$ | $t(chat|cat)$ | $t(chat|the)$ |
|----------------|--------------|--------------|
| $1/3$          | $1/3$        | $1/3$        |

| $t(le|black)$ | $t(le|cat)$ | $t(le|the)$ |
|----------------|--------------|--------------|
| $1/3$          | $1/3$        | $1/3$        |

| $t(noir|black)$ | $t(noir|cat)$ | $t(noir|the)$ |
|----------------|--------------|--------------|
| $1/3$          | $1/3$        | $0$          |

Do the Expectation step: first compute the probability of each possible alignment of each sentence:

- “black cat”: “chat noir”
- “the cat”: “le chat”

Normalise the fractional ‘counts’ for each pair and each source:

| Alignment | P(a|f,e) |
|-----------|---------|
| black     | $1/2$   |
| chat      | $1/2$   |
| noir      | $1/2$   |
| le        | $1/4$   |
| the       | $1/2$   |
| chat      | $1/2$   |
| noir      | $0$     |

Finally sum the fractional ‘counts’ for each pair and each source:

<table>
<thead>
<tr>
<th>Pair</th>
<th>Total Fractional Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>black/chat</td>
<td>$1/2$</td>
</tr>
<tr>
<td>black/noir</td>
<td>$1/2$</td>
</tr>
<tr>
<td>black/the</td>
<td>$1/2$</td>
</tr>
<tr>
<td>chat/noir</td>
<td>$1/4$</td>
</tr>
<tr>
<td>chat/the</td>
<td>$1/2$</td>
</tr>
<tr>
<td>noir/the</td>
<td>$0$</td>
</tr>
</tbody>
</table>

The maximisation step: normalise the counts to give ML estimates:

| Alignment | P(a|f,e) |
|-----------|---------|
| black     | $1/2$   |
| chat      | $1/2$   |
| noir      | $1/2$   |
| le        | $1/4$   |
| the       | $1/2$   |
| chat      | $1/2$   |
| noir      | $0$     |

All the correct mappings have increased, and some of the incorrect ones have decreased.

Feeding the new probabilities back in, what we now see for each alignment is:

| Alignment | P(a|f,e) |
|-----------|---------|
| black     | $1/2$   |
| chat      | $1/2$   |
| noir      | $1/2$   |
| le        | $1/4$   |
| the       | $1/2$   |
| chat      | $1/2$   |
| noir      | $0$     |

And the right answers have pulled ahead.