Multilingual and Parallel Corpora

Machine Translation

Amir Zeldes

amir.zeldes@georgetown.edu

Plan

- Some history
- Corpora and machine translation:
  - Corpora as example banks
  - Rule based approaches
  - Statistical approaches
  - Neural MT (very superficially)

Disclaimer

- This is not a Machine Translation course!
- If you're interested in learning more about MT, consider taking the Statistical Machine Translation class next year (over-under, Spring semester)
  - Not heavy on programming, but some experience recommended
  - Possible preparation for those more seriously interested: take Intro to NLP (LING-362) in the Fall

Some history

- Machine translation began in the 40s, shortly after early computers were designed to break German and Japanese codes (Enigma, Ultra…)
- Partly viewed as a problem of cryptography:
  - *When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode'. [Warren Weaver, 1947]*

Warren's Memorandum

- Published in 1949, the memo suggested some directions for research that resonated with policy makers and funding agencies
- Four key points:
  - Ambiguity could be solved via *n*-word context (fast = rapid/motionless); hypothesis: *n* is small
  - Translation is akin to logical theorem proof; if the ST conveys some information, TT must be provable from it
  - The cryptography postulate: language choice is really a cryptographic code
  - Universal grammar should make MT possible

Warren's Memorandum

- *Think, by analogy, of individuals living in **a series of tall closed towers**, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, **when an individual goes down his tower**, he finds himself in **a great open basement**, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers*

The IBM-Georgetown Experiment

- Research along Warren's lines of reasoning led to substantial funding culminating in the 1954 experiment at Georgetown
- The experiment tested MT using only:
  - A lexicon of 250 items
  - 6 rule types
  - 60 sentence pairs
- A public presentation of the system was very widely publicized and led to statements that MT would be a solved problem in a matter of years

The IBM-Georgetown Experiment

- Examples

Russian original	English translation
My pyeryedayem mislyi posryedstvom ryechyi.	We transmit thoughts by means of speech.
Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu.	Magnitude of angle is determined by the relation of length of arc to radius.
Myezhdunarodnoye ponyimanyiye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyityichyeskix voprosov.	International understanding constitutes an important factor in decision of political questions.

Rule types

- Operation 0 – exact equivalent for a translated item
- Operation 1 – rearrangement AB -> BA
- Operation 2 – multiple translations, look ahead: selection based on max 3 subsequent words
- Operation 3 – multiple translations, look behind (max 3)
- Operation 4 – Omission of lexical/morphological item (null alignment)
- Operation 5 – Insertion of the lexical/morphological item (null fertility)

➢ NB: No operation exceeds sentence context!

Criticism – Bar Hillel's *pen* example

- What would you need to know to translate this:

The box was in the pen

"Whenever I offered this argument to one of my colleagues working on MT, their first reaction was: "But why not envisage a system which will put this knowledge at the disposal of the translation machine?" Understandable as this reaction is, it is very easy to show its futility. What such a suggestion amounts to, if taken seriously, is the requirement that a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion." (Bar Hillel 1960)

Please read for next time!

The ALPAC report

- Progress in the late 50s was slow
- By the mid 60s, funding agencies wanted an account
  - Automatic Language Processing Advisory Committee (ALPAC)
  - 1966 report showed little progress despite vast investment
  - Annual investment in translating Russian manually was only $20,000,000, suggesting MT was a waste of time
  - Little interest in massive translation (e.g. to follow Russian scientific publications, literature)

Where do we get this knowledge from?

- Early attempts to incorporate increasing amounts of encyclopedic knowledge in MT systems were largely unsuccessful
- But then came the corpora…
  - (Lots of) text as a proxy for world knowledge
  - How to generalize to unseen cases?
  - Where does this fall short?

The corpus as an example bank

- Parallel corpora can be seen as collections of translations
  - Thinking about unit size
  - Is discourse coherence necessary? -> Wikipedia sentence pairs
- Can we make an **algorithm** to translate an input A?
  - Look for A in the corpus
  - Retrieve all aligned units: B, C, D…

The corpus as an example bank

- What can B, C, D be?
- This depends on alignment granularity:
  - A text (not very useful)
  - A paragraph
  - A sentence (most often)
  - A phrase
  - A word (also think about bilingual dictionaries)

Translation Memory

- TMs are systems used by translators to find **examples** of translations for specific search terms
- Translations are saved and indexed on multiple levels
  - Sentences
  - Phrases
  - Words
- One of the first commercially successful applications of digital parallel corpora (long before MT was any good!)

The corpus as an example bank

- What can B, C, D be?
- This depends on alignment granularity:
  - A text (not very useful)
  - A paragraph
  - A sentence (most often)
    - Easiest to align, hard to find exact attestation
  - A phrase –
    - Somewhat frequent, less ambiguous
  - A word –
    - highest chance to find instances
    - Ambiguous
    - Alignment inaccurate

Alignment below the sentence level

- The entire sentence we want to translate will only rarely appear verbatim
- Differences can sometimes be small
  - e.g. add/remove an adverb
  - optional word order variation (scrambling)
- Sometimes entire phrases appear, but in meaningfully different constellations

Search and result

- Input: “not to mention other problems”
- Target: German
- Found in the corpus:

Der Wettbewerb , d. h. der Krieg , den die großen Unternehmen untereinander austragen , schlägt sich ständig in Entlassungen und Betriebsschließungen nieder , **ganz zu schweigen** von der riesigen Verschwendung produktiver Kapazitäten .

Competition is a war which has major concerns fighting each other , which constantly takes the form of layoffs , factory closures , **not to mention** extensive waste of production capacity .

Search and result

- Part of what we are searching for does not appear
- How can we tell what parts correspond to our search terms?

Der Wettbewerb , d. h. der Krieg , den die großen Unternehmen untereinander austragen , schlägt sich ständig in Entlassungen und Betriebsschließungen nieder , ganz zu schweigen von der riesigen Verschwendung produktiver Kapazitäten .

Competition is a war which has major concerns fighting each other , which constantly takes the form of layoffs , factory closures , **not to mention** extensive waste of production capacity .

What we would like

- Alignment on **all** levels (what corresponds exactly to “not to mention”?)
- A mechanism that can **combine** translations:
  - One source delivers: “ganz zu schweigen von”
  - Another tells us: “other problems” = “die anderen Probleme”
  - The mechanism should find the best match (if not “ganz zu schweigen” then “ganz” + “zu schweigen” + …)

What we would like

- Alignment on **all** levels (what corresponds exactly to "not to mention"?)
- A mechanism that can **combine** translations:
  - So we get: "ganz zu schweigen von die anderen Probleme"
  - Good?
  - "ganz zu schweigen von den anderen Problemen"

What we would like

- Alignment on **all** levels (what corresponds exactly to "not to mention"?)
- A mechanism that can **combine** translations:
  - One source delivers: "ganz zu schweigen von"
  - Another tells us: "other problems" = "die anderen Probleme"
  - And another issue - how to prioritize fallbacks:
    - If you can't find "not to mention" maybe "not" + "to mention" + …)
    - Good?
    - Why not: "not to" + "mention"….?

How much context?

- What kind of background information does this mechanism need to make such decisions?
- What types of computationally extractable annotations?
  - Part of speech, case …
  - Maybe we can fix *von + dative*?

- Try translating this sentence to your L2:
  - *It was running for a seemingly endless time until finally it came to a stop.*

How about now?

- The dog’s heart sank faster than it had just fallen and knew that it had to escape quickly. It was running for a seemingly endless time until finally it came to a stop.

And now?

- Aunt Clara was doing her best not to cry, but now a single tear, threatening to drop from the sad eye directed at them, presently decided to trickle away after all. It was running for a seemingly endless time until finally it came to a stop.

And now?

- He knew he couldn’t let anyone know he’d gotten sick, but his damned nose was going to give him away. It was running for a seemingly endless time until finally it came to a stop.

What do humans do?

- A human translator:
  - Reads the text in the SL
  - **Understands** it
  - Produces a (the?) translation in the TL

- Is there an algorithm?

Rule based approaches to MT

- We could make a (large) list of rules:
  - “not to mention” + NP >
  - “ganz zu Schweigen von” + NP >

- If the language pair behaves regularly we can **map** their words and even word orders (a.k.a. **transfer**)
- Would we need to understand the content?

Example Japanese - Chinese

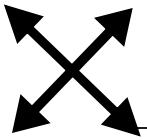
- Chinese is SVO, Japanese is SOV
- Japanese is agglutinative, Chinese is isolating
- Both languages use an ideographic script (in Japanese also two syllabaries)
- Often symbols are identical for the same concepts

➢For simple cases, can we just use symbol reordering?

Example Japanese - Chinese

- Simple declarative transitive sentence:

S	O	V
Student	English	learns
大学生は	英語を	学ぶ
大学生	学	英語
Student	learns	English
S	V	O



Example Japanese - Chinese

- Simple declarative transitive sentence:

S	O	V
Student	English	learns
大学生は	英語を	学ぶ
大学生	学	英語
Student	learns	English
S	V	O