Multiple reference translations for European languages

Christian Buck, Daniel Zeman, Eva Hasler

Idea

- Create multiple reference translations by paraphrasing a single reference
- Use Pivot approach to create paraphrase table
- Train **en-en** paraphrasing system
- Tune paraphrasing system on multiple English references from another data set

Pivot approach to paraphrasing

- go forward and backward in phrase table
- for each English phrase find all French translations
- for each French translation find all English translations
- done!

(Bannard & Callison-Burch 2005)

Example

take

⇨ nehmen, einnehmen

⇨ take, consume, conquer, take away, eat

Filtering of paraphrases

Extraction generates *lots* of rubbish

Filter using:

- Heuristics (e.g. no substrings)
  - better: remove substrings only if additional subphrase unaligned (keep "though"/"although")
- Scores (e.g. keep only top-20)
- Absolute phrase length (e.g. max. 4)
- Consistency across different pivot languages
- Consistency across different topics

Filtering continued

- Consider quality of induced alignments

  f1 f2 f3 ||| e1 e2 e3 ||| … ||| 1-1 **2**-2
  f1 f2 f3 ||| e3 e4 ||| … ||| 0-0 0-1 **2**-1

  e1 e2 e3 ||| e3 e4 ||| … ||| **2-1**

  Ignore paraphrase if
  - aligned token is punctuation
  - alignment missing altogether

Data/models

- TED talks corpus (collection of talks)
- language pairs: fr-en, de-en
- HTMM, filter using stop words, 10 topics, “unknown” topic if filtering left empty sentences
- compute topic model on one side of the corpus → map topics to other side
- translation should preserve the topic, though lexical differences between languages might induce different topic models

"economy" ||| einnehmen ||| earn ||| ...
"economy" ||| einnehmen ||| to make money ||| ...

→ earn ||| to make money ||| ...

"medicine" ||| einnehmen ||| to take orally ||| ...
"medicine" ||| einnehmen ||| to ingest ||| ...

→ to take orally ||| to ingest ||| ...

"politics" ||| einnehmen ||| to conquer ||| ...
"politics" ||| einnehmen ||| to occupy ||| ...

→ to conquer ||| to occupy ||| ...

Paraphrase tables

- Identity feature: train a weight to learn how often we should translate a phrase into itself vs. into a paraphrase
- avoid including topic features into phrase table by splitting corpus according to topics
- extract separate sets of paraphrases on these subsets of the training data, merge tables
  → valid set of paraphrases
- extract paraphrases from entire training data and filter with merged paraphrase table

Examples of extracted paraphrases

1.3 million ||| one point three million ||| ...
good!

10 miles ||| 15 kilometers ||| ...
maybe?

10 percent ||| 10 mm |||...
10 percent ||| 30 ||| ...
10 percent ||| make 10 ||| ...
bad..

220 ||| about 110 ||| .... **aaaahrr!**

Examples of extracted paraphrases

pretty much ||| pretty much ||| ...
pretty much ||| pretty sure ||| ...
pretty much ||| quite different ||| ...
pretty much ||| quite ||| ...
pretty much ||| rather ||| ...
pretty much ||| real ||| ...
pretty much ||| really , really ||| ...
pretty much ||| really ||| ...
pretty much ||| relatively ||| ...
pretty much ||| see very ||| ...

TODO

- Train en-en translation systems using paraphrase table, tune on sets of multiple English references (NIST data)
- Translate references!
- **Cross language extraction**
  - Intersect paraphrase tables resulting from the two different language pairs
- Compare both paraphrasing approaches
- **Evaluate**
  - use multiple references in tuning step