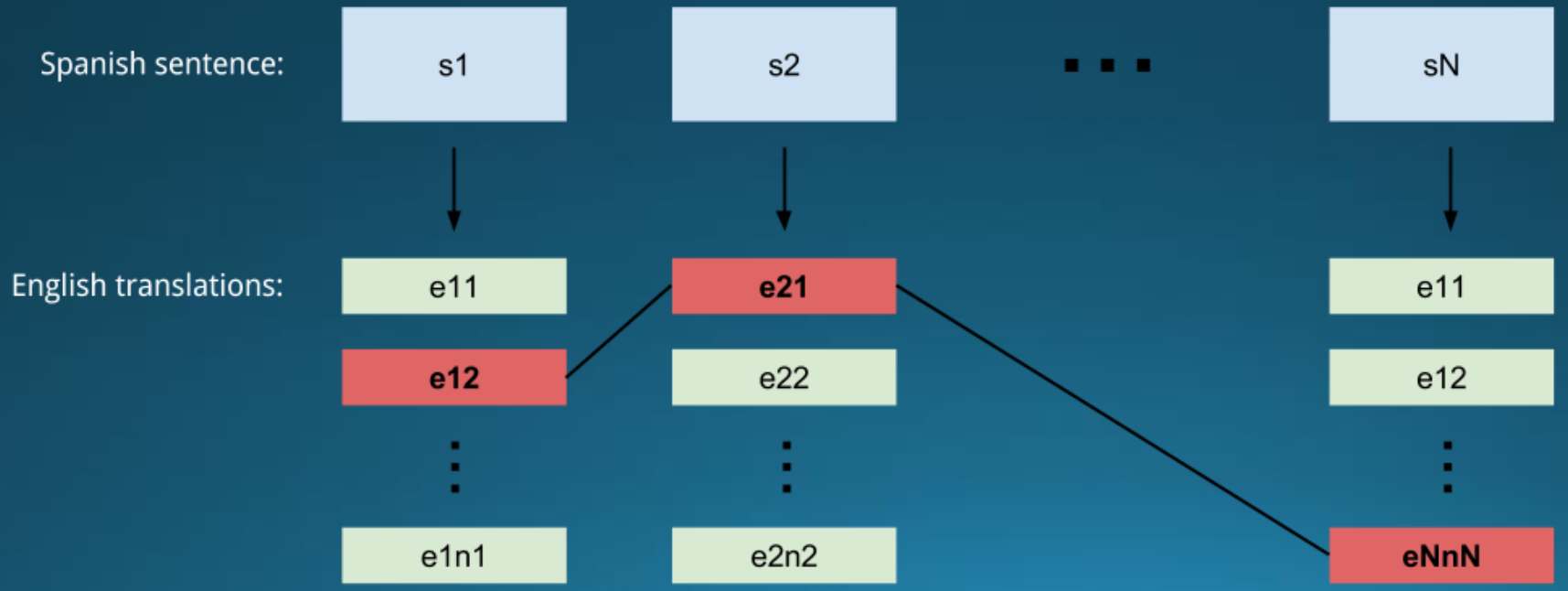
Jussi Kolehmainen 24.10.2013

Statistical Machine Translator

The Problem

- Translate a Spanish sentence into English
- Translate word by word
  - →Multiple translations for each word



One approach

1. Build index from a large corpus (e.g. frequencies)
2. Translate word by word with a dictionary
3. Calculate statistical features from the index and estimate probabilities for the alternative translations
4. Rank the alternatives by their probabilities

Corpus → Index

- European Parliament Proceedings Parallel Corpus 1996-2011
- 1,965,734 sentences in both Spanish and English

- Redis NoSQL database. Example keys:
  - en:word:occur → 3452 (frequency)
  - en:stem:occur → [”occurs”, ”occurence”, …] (different forms)
  - en:word:occur:sentences → [523, 1523, 6534, …] (sentences ids)
  - en:bigram:has:occurred → 125 (frequency)

Dictionary API

- Glosbe API
  - *http://glosbe.com/gapi/translate?from=**spa**&dest=**eng**&format=**json**&phrase=**hola***
- Response has both direct translations and meanings
  1. Direct translations
     - "hola" → "hello"
     - Candidates for translations
  2. Meanings
     - "killed" → "… form of **kill**" → New query with "kill"

Probability models

1. Language model P1

$$P(e_2, e_1) = \frac{\#(e_1, e_2)}{\#(e_1)}$$

2. Translation model P2

$$P(e|s) = \frac{\&(e, s)}{\#(e)}$$

= frequency in the same sentence
& = frequency in the parallel sentences

Ranking

- Probabilities are smoothed
  = some part of the probability to non-existent bigrams and translations
- Probability for a translation s → e: P = P1 * P2
- Initial probabilities by word frequencies
- Probability estimation is iterated N times
- Alternatives are ranked by their probabilities

Conclusions

- Problems
  - Dictionary API has a limit for queries per IP address
  - Dictionary API is relatively slow → Switch to offline dictionary
  - Finding a good indexing method for tens of millions of entries

- Some results
  - Redis database is fast and east-to-use for indexing sentences
  - Simple statements produce meaningful translations
  - Corpus topic affects the results