Statistical Machine Translation

Cristina Vertan

University of Hamburg • Informatics Department

Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/

E-Mail: vertan@informatik.uni-hamburg.de

Statistical MT-Principles - 1 -

- Given:
  - A source sentence (e.g in German.): D = d_1, ...d_i,...., d_n (d_i are the words) which has to be translated into a sentence (in English for e.g..) E = e_1, ...e_i,...., e_m .
  - A parallel aligned german-english corpus

Between all translation possibilities it is searched the one with the highest probability.

- This means mathematically :

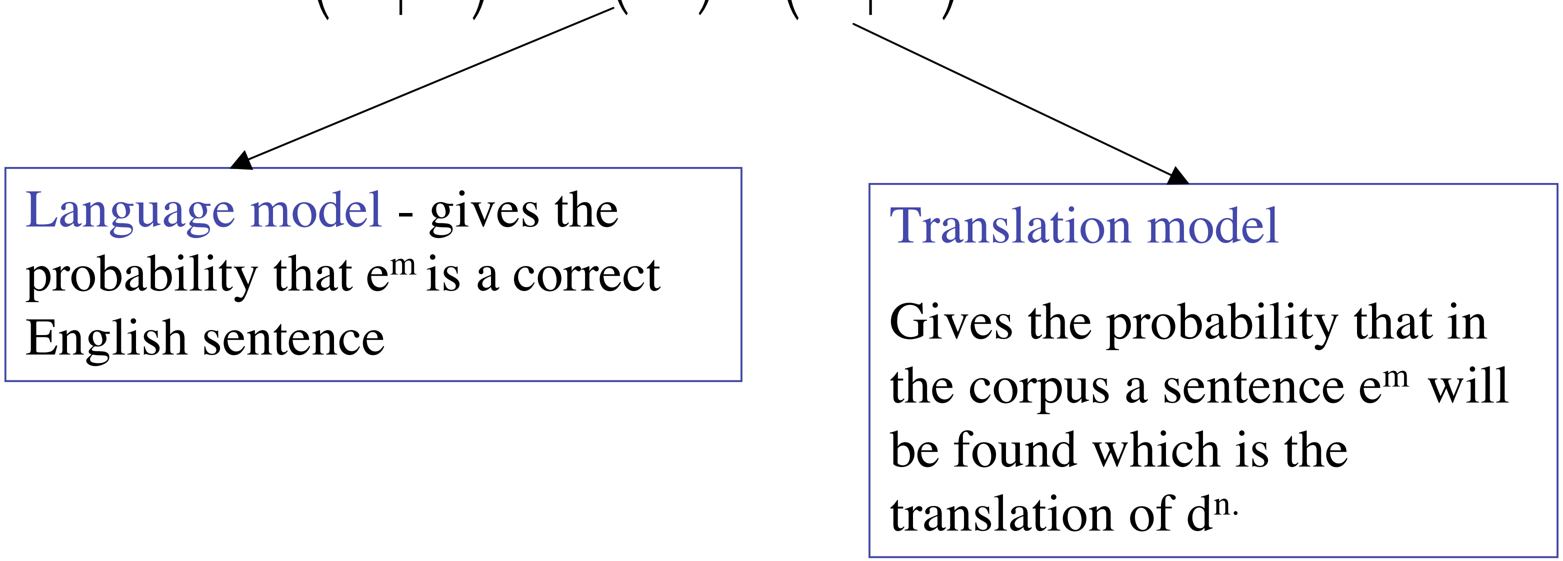
$$\hat{e}^m = \arg\max_{e^m} \left\{ \Pr\left(e^m \middle| d^n \right) \right\}$$

- Depending on how this probability is calculated there are different models for the translation.

Statistical MT -Model 1

- Das Source-Channel Modell (used very often):
  - Following decomposition is used:

$$\Pr\left(e^m \middle| d^n\right) = \Pr\left(e^m\right) \Pr\left(d^n \middle| e^m\right)$$



Language model - gives the probability that e^m is a correct English sentence

Translation model

Gives the probability that in the corpus a sentence e^m will be found which is the translation of d^n.

Both models are dependent of parameters, which are calculated in the training phase

Statistical MT- Model 2-

- Direct Maximum Entropy Translation Model
  - The original probability is calculated directly, following different translation features (mathematically is a function with parameters)

$$\Pr\left(e^{m}\middle|d^{n}\right)$$

- Alignment Model
  - A new parameter is introduced, which models the alignment-mapping. Here features like Fertility and Distortion are considered

Fertility, Distorsion -Reminder

Die $_{1}$ Oppositionsfraktionen $_{2}$ im $_{3}$ baden - wuerttembergischen $_{4}$ Landtag $_{5}$ haben scharfe $_{6}$ Kritik $_{7}$ an der Finanzpolitik $_{8}$ der $_{9}$ CDU / FDP - $_{10}$ Koalition $_{11}$ geuebt $_{12}$.

The (1) opposition parties (2) in (3) Baden - Wurttemberg 's(4) Landtag(5) have strongly (6) criticized(7+12) the financial policies(8) of (9) the governing CDU / FDP (10) coalition(11) .

Fertitlity of a source word = the number of words in the target text

e.g. *fertility(Oppositionsfraktionen)* = *2*

Distortion = Source and target words do not appear in the same place e.g. Koalition und coalition

Example

Advantages of Statistical MT

- Use no linguistic knowledge (as long as the alignment of the corpus is done automatically)

- Loose dependencies between constituents can be modelled better with statistical models as with rules

- It is especially indicated to be used in embedded systems e.g. in Speech Systems, where a language model already is defined (for the speech recognizer)

Well-known problems with Statistical MT

- New field, there are few systems which can be evaluated. (Verbmobil, Translation of Canadian parliament debates)
- Exceptions can be trained difficult
- Morphology:
  - Inflected forms of the same word are treated as not-related words. E.g the Word *diriger* in French is translated with *führen* or *leiten in German*. For each one of the 39 inflected forms of the word the model has to be trained (which is translated with *führen* and which with *leiten*)muss).
- Not-local dependencies are difficult to be trained. The System produces usually correct word-translations but in an incorrect order
- Probabilities for rare words are not to be trusted.
- The models are very sensible to data-changes.

Example of incorrect Translations with statistical MT -1-

- Source text: *Permettez que je donne un example à la chambre.*
- Correct translation: *Let me give the House an example*
- System-Translation: *Let me give an example in the House*

The model gives the highest probaility for the word à the word in.

© Manning und Schütze S. 489

Examples of incorrect translations with statistical MT -2-

- Source text: *Vous avez besoin de toute l'aide disponible*
- Correct Translation: *You need all the help you can get*
- System-Translation: *You need of the whole benefits available.*

In the aligned corpus the translation of toute l'aide is found with whole benefits and simply used further

© Manning und Schütze S. 489

Comparison of linguistic and empirical methods -1-

- In Verbmobil-System (German-English-Japanese Speech-to-Speech System) were 3 MT approaches implemented, 1 transfer-based, 1 statistical and 1 example based.
- After the evaluation (number of sentences evaluated as not correct translated)
  - Semantic Transfer 62 %
  - Example based MT 35%
  - Statistical MT 29%

Comparison of linguistic and empirical methods - 2-

- Most problems of the empirical approaches are due to:
  - Word order: the target language model is not trained accordingly
  - Disambiguation: Very difficult for prepositions which are translated according to the context
  - No partial translation: statistical translation process only sentences. If the speech recognizer contains mistakes or „false starts“ the quality of the translation decrease rapidly.
  - Problems with verb particles: in this case is a morphological preprocessing needed, but this increases very much the processing time

Comparison of lingusitic and empirical methods -Examples 1-

- Spoken: Okay. Dann treffen wir uns um acht Uhr sechs auf dem Hauptbahnhof in Hamburg. Ist das okay?
- Recognized: okay dann treffen wir uns um acht Uhr sechs auf dem Hauptbahnhof in hamburg. Ist das okay
- Transfer: okay | that is okay
- Example: and then we will meet in Hamburg at eight six is that okay
- Statistical: okay, then we will meet at eight oh six at the main train station, in hamburg, is that okay.

Comparison of lingusitic and empirical methods - Examples 2-

- Spoken: Dann muss ich Ihnen gleich mal sagen, die ersten zwei Oktoberwochen sieht‘s bei mir ganz schlecht aus.
- Recognized: dann muss ich Ihnen gleich mal sagen die erste zwei Oktober-Wochen sieht‘s bei mir ganz schlecht aus sind
- Transfer: -
- Example: then I have to tell you right away it looks quite bad on my end in on the first two weeks of October I
- Statistiscal: then I have to tell you right away, the first two weeks of October is very bad for me, is

Comparison of lingusitic and empirical methods -Examples 3-

- Spoken: Das tut mir leid, da habe ich - , muss ich auf eine Messe.
- Recognized: das tut mir leid da ich mus ich eine Messe
- Transfer: I‘m sorry about that I must a fair there
- Example: -
- Statistical: I am sorry, I have got a fair.