Example-based Machine translation with linguistic knowledge

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

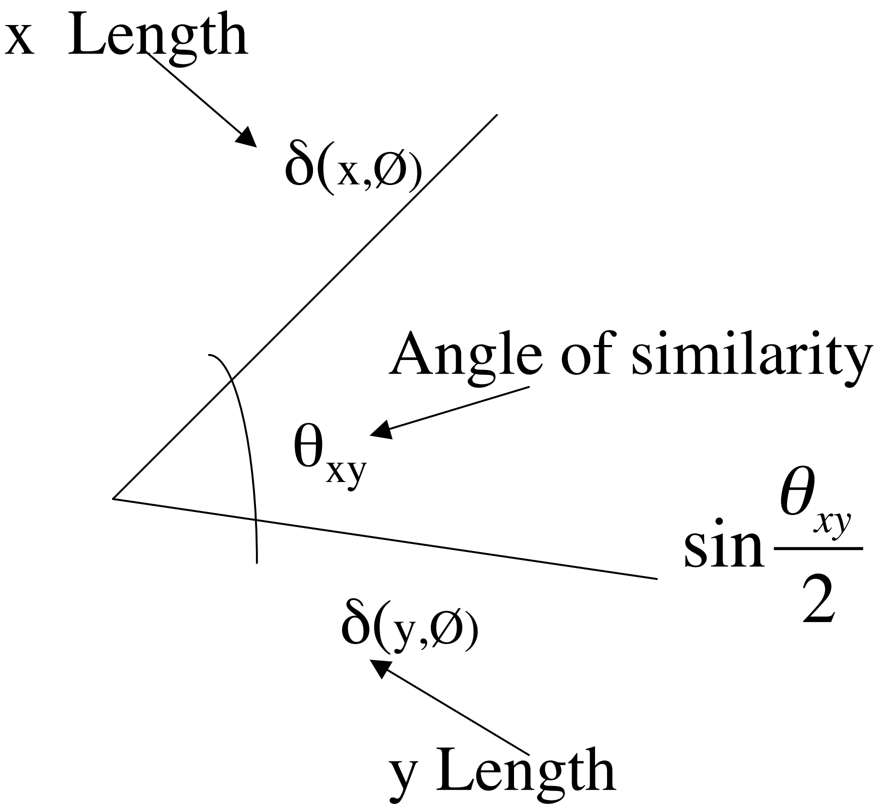
EBMT with morphological/lexical knowledge

- Use only the stems when measuring the distance between input and entries in the database
- Mark in the database words with unabigous function (e.g conjunctions)
- Whenever possible allign fixed expressions
- When measuring the Edit distance look at the PoS of the words

Word-based Matching: ”Angle of similarity” - 1 -

- A trigonometrical distance is computed.
- The distance between 2 sentences corresponds to a difference function δ.
- This difference function works similar as the string-based matching (the number of operations is calculated)
- The operations are weighted, e.g. the insertion of a comma has a smaller weight than the absence of an adjective.
- The weights are defined according to the system and the translation domain

Word-based Matching: ”Angle of similarity” - 2 -



Distance between sentence x and sentence y

$$\sin\frac{\theta_{xy}}{2} = \frac{\delta(x,y) - \left|\delta(x,\varnothing) - \delta(y,\varnothing)\right|}{2\times\min\left\{\delta(x,\varnothing),\delta(y,\varnothing)\right\}}$$

Word-based Matching - ”Angle of similarity” Example

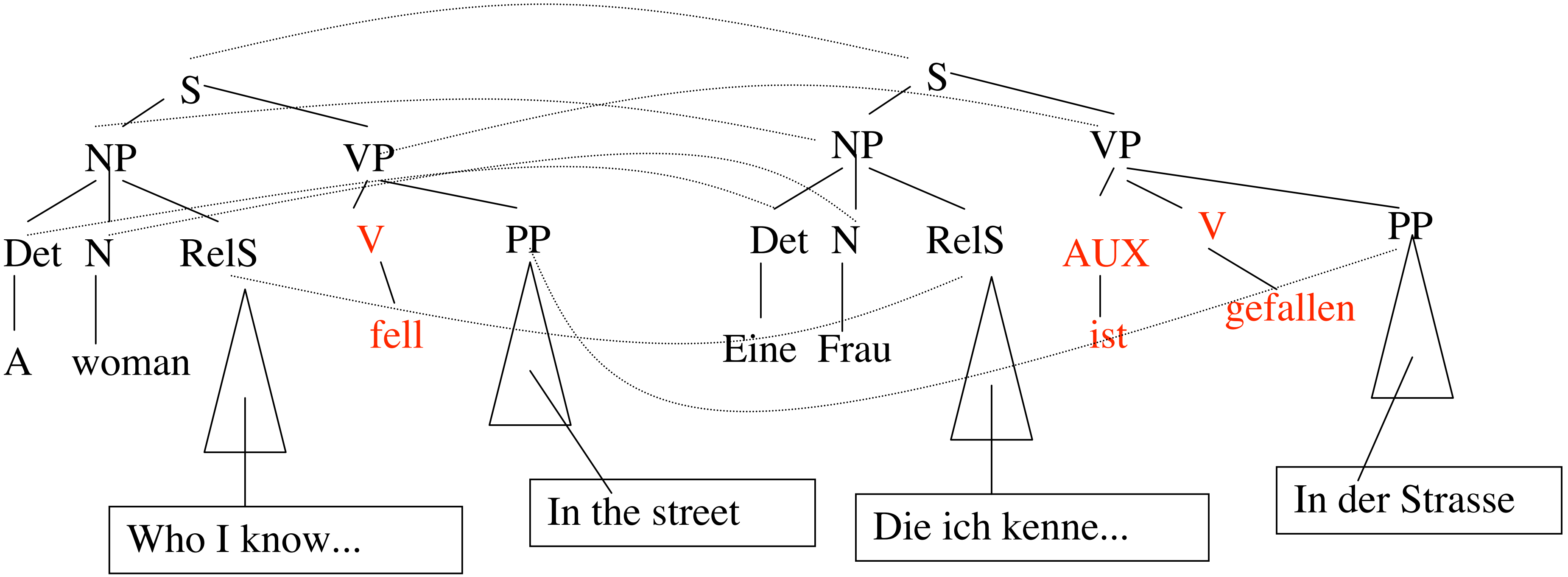
1. *Lesen Sie Seite 3 im Kapitel “Benzin”*
2. *Lesen Sie Seite 3 im Kapitel “Benzin” und Seite 5 in Kapitel “Länderspezifische Bemerkungen”*
3. *Lesen Sie Seite 4 im Kapitel “Bremsen”.*

- String-based matching gives a closer similarity between sentence 1 and sentence 3 because they differ only by 1 word.

  However: Sentence 2 is actually a better choice as sentence 1 is contained entirely. This choice is made by the “angle distance”.

EBMT with syntactic knowledge

- The Translation patterns are not words , but syntactical structures in both languages with corresponding links



What to do in the Practical Exercice ?

User Interface Group

- Design a GUI where:
  - Text is typed in
  - The language pair for translation can be set and as consequence the corresponding resources are loaded
  - Translation is showed
- Define the communication methods and formats with the database(s) and lexicon(s) (together with the language resource groups)
- Define the communication format with the matching and recombination processes (together with the matching and recombination groups)

Language Ressource Groups -1-

- Separate lexicon from the examples database.
- Try to cover in the bilibgual lexicon as much as possible from the domain. In the lexicon indicate at least the Stem and a pointer to the corresponding word in the other language

```
<entry  id=„ro_003“ tr=„en_005“>
     <word> vizitează </word>
    <stem> vizita </stem>
</entry>
```

Language Resource Groups -2-

- Try to cover in the database a variety of structures: NPs, PPs, VPs, but also complete sentences.
- For syntactic structures that you feel that repeat try to include some syntactic patterns
- Organise optimal your database, e.g make an index: for each word in your database list all sentences in which appear. This could help the faster search in the database.

Matching Groups

- Perform „Edit Distance“, respectively „Angle of Similarity“ to detect the candidate most similar phrases in the database (you have to impose a certain treshold)
- Each of the selected phrases by the edit distance compare with the input and extract the longest common sequence
- Send these sequences to the recombination module
- Perform the operations first without and then with morphological information

Recombination Groups

- Try to identify , for the retrieved chunks in the SL their correspondents in TL
- Try to resolve parts with overlap.
- You have to obtain from the matching groups not only the chunks in the SL but also the order in which they appear in the input.
- Try to put together the correspondent chunks, first without any other information. Then define some simple combination rules
- Try to use the syntactic information in the database