CS 294-5: Statistical Natural Language Processing

Machine Translation

Dan Klein

includes slides from Manning (from Yamada, Knight)

Machine Translation

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

REF: According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959 billion US dollars of foreign capital, including 40.007 billion US dollars of direct investment from foreign businessmen.

IBM4: the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

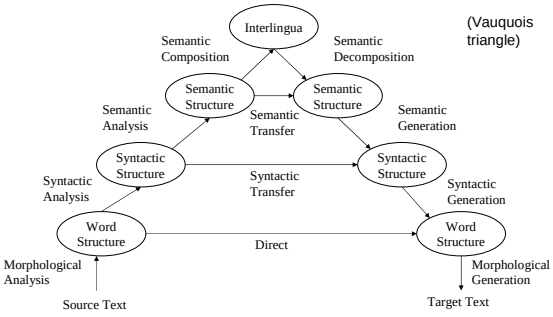
Yamada/Knight: today's available data of the Ministry of Foreign Trade and Economic Cooperation shows that china's actual utilization of November this year will include 40.007 billion US dollars for the foreign direct investment among 46.959 billion US dollars in foreign capital

History

- 1950's: Intensive research activity in MT
- 1960's: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: `Recovery period'
- 1975-1985: Resurgence (Europe, Japan)
- 1985-present: Gradual Resurgence (US)

http://ourworld.compuserve.com/homepages/WJHutchins/MTS-93.htm

Approaches



Just a Code?

- "Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in 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 (1955:18, quoting a letter he wrote in 1947)

Bag Generation

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

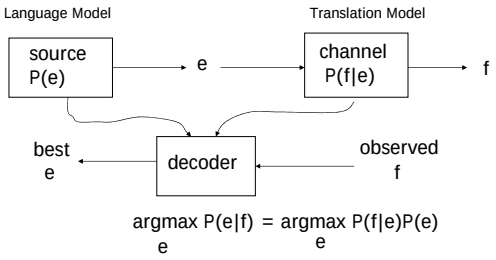
Reconstruction preserving meaning (8 of 38)

Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

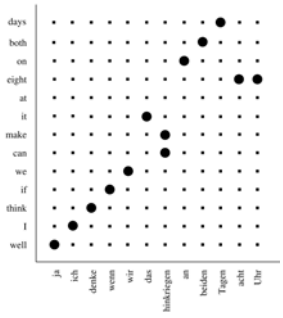
In our organization research has two missions.
⇒ In our missions research organization has two.

MT System Components

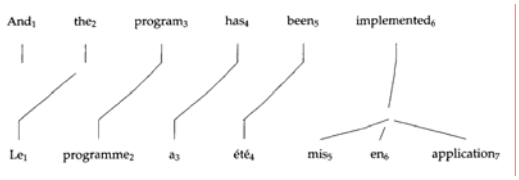


$$\text{argmax}_e P(e|f) = \text{argmax}_e P(f|e)P(e)$$

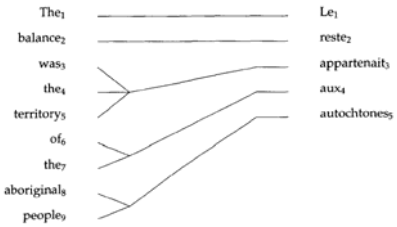
Word-to-Word Alignment



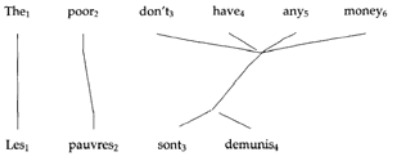
Word-to-Word Alignment (1-Many)



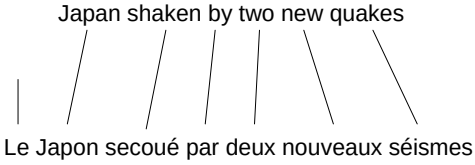
Word-to-Word Alignment (Many-1)



Word-to-Word (Many-Many)



Monotonic Translation



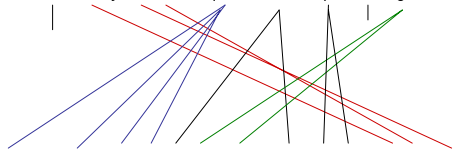
Order Change

Japan is at the junction of four tectonic plates

Le Japon est au confluent de quatre plaques tectoniques

Phrase Movement

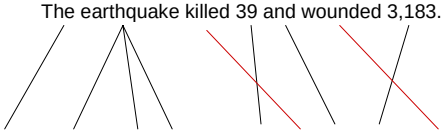
On Tuesday Nov. 4, earthquakes rocked Japan once again



Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.

Head Switching

The earthquake killed 39 and wounded 3,183.



Le tremblement de terre a fait 39 morts et 3,183 blessés.

Non-Literal Translation

Injuries were also avoided by the automatic shutdown of a train.

Un train s'est également arrêté sans qu'aucun passager ne soit blessé.

IBM Model 1

- Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes

Learning with EM

- Model 1 Parameters: P(f|e)
- Start with P(f|e) uniform, including P(f|null)
- For each sentence:
  - For each French position i
    - Calculate posterior over English positions, P(a_i|i)

$$P(a_i \mid i) = \frac{P(f_i \mid e_{a_i})}{\sum_{a_i'} P(f_i \mid e_{a_i'})}$$

    - Increment count of word f_i with word e_{ai}
- Iterate until convergence

IBM Model 2

HMM Alignment Model

Modeling Fertility

Cascaded Training