The Machine Translation Research of Xiamen University and its Partners

Tangqiu Li

Department of Computer Science Xiamen University

1. Introduction

- Our views of the state of the art of machine translation
- underlined methodology and basic structure of our systems
- The hybrid method for syntactic and semantic structural disambiguation of Chinese.
- An introduction of the work of ours and our partners(The Team of Prof. Huowang Chen and Dr. Xiaodong Shi)
- Our directions and plans of our future research and development.

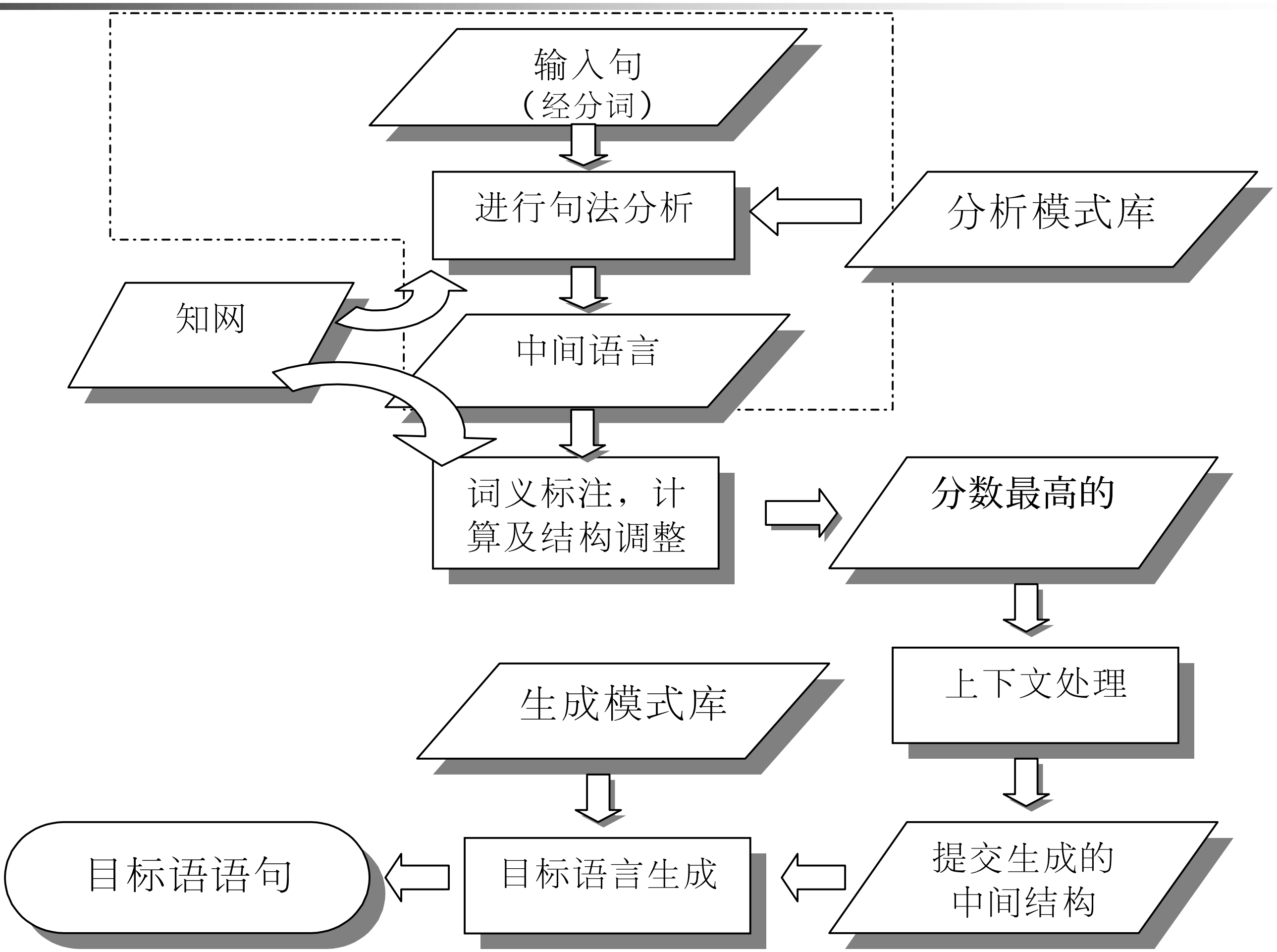
2. Our Vision About The state of Art of MT

- MT has been studied for more than 50 years, but a general, full automatic MT with high quality is still a dream not a reality.
- A lot of method has been developed to solve the problem:Rule-Based, Example Based or Statictics based. However they are only useful means complement each other to approach the goal, none of them can lead to the goal in isolation.
- The problem is that NL is the tool of intelligent human beings. Translation of language needs understanding.
- Knowledge representation and application is a core issue. But humans understand so little about themselves, that the problem has not been well solved yet in AI.
- Before the theoretical and technical breakthrough about the knowledge representation and application, we can still do a lot of things to approach the goal, and make it closer and closer.

3. Our Current Tactics With MT

- Translation based on interlingua
- In the believe that humans have a common structure in the concept level, the interlingua is defined as close to the concept frame representation as possible.
- Emphasize on the usage of knowledge in the processing, whether it is from the knowledge base built by human expert or one from language corpus.
- A hybrid method for sentence structure and semantic disambiguation using both rule based method and statistics based one.
- Reasoning about missing information in a sigle sentence through Local discourse context.

4. The Structure of Our MT System



5. The problem of Disambiguation

- There are two major sources of ambiguity in language analysis. One is lexical (polysemy). The other is structural, both of which lead to semantic ambiguity.
- For example, the structures of the following two phrases look the same, but their proper syntactic structures are much different.
  A. “维修图书馆的空调”
  v　　n　　的　　n
  B. “装修图书馆的工人”
- A.“(维修(图书馆　的　空调))　”
  B.“((装修　图书馆)的　工人)”
- Disambiguation soly depending on the syntax does not work well more ofter than not.

How to deal with…(traditional)

- To solve such kind of problem, one method is to appeal to world knowledge or **semantic constraints**:
  1.“空调(air conditioner)” could not act as the subject of “维修(repair)”.
  2.“工人(worker)” could not act as the object of “装修(decorate)”.
  Semantic constraints can be useful, but… (Semantic constraint acquiring and encoding is a very difficult task)
- The other method is **Example-based**. It finds the most similar example with input sentence from a large-scale bilingual corpus, and then adjust the counterpart of object language as the translation. It voids the manual knowledge acquiring and encoding, but building a properly aligned large scale bilingual corpus is also a major undertaking.

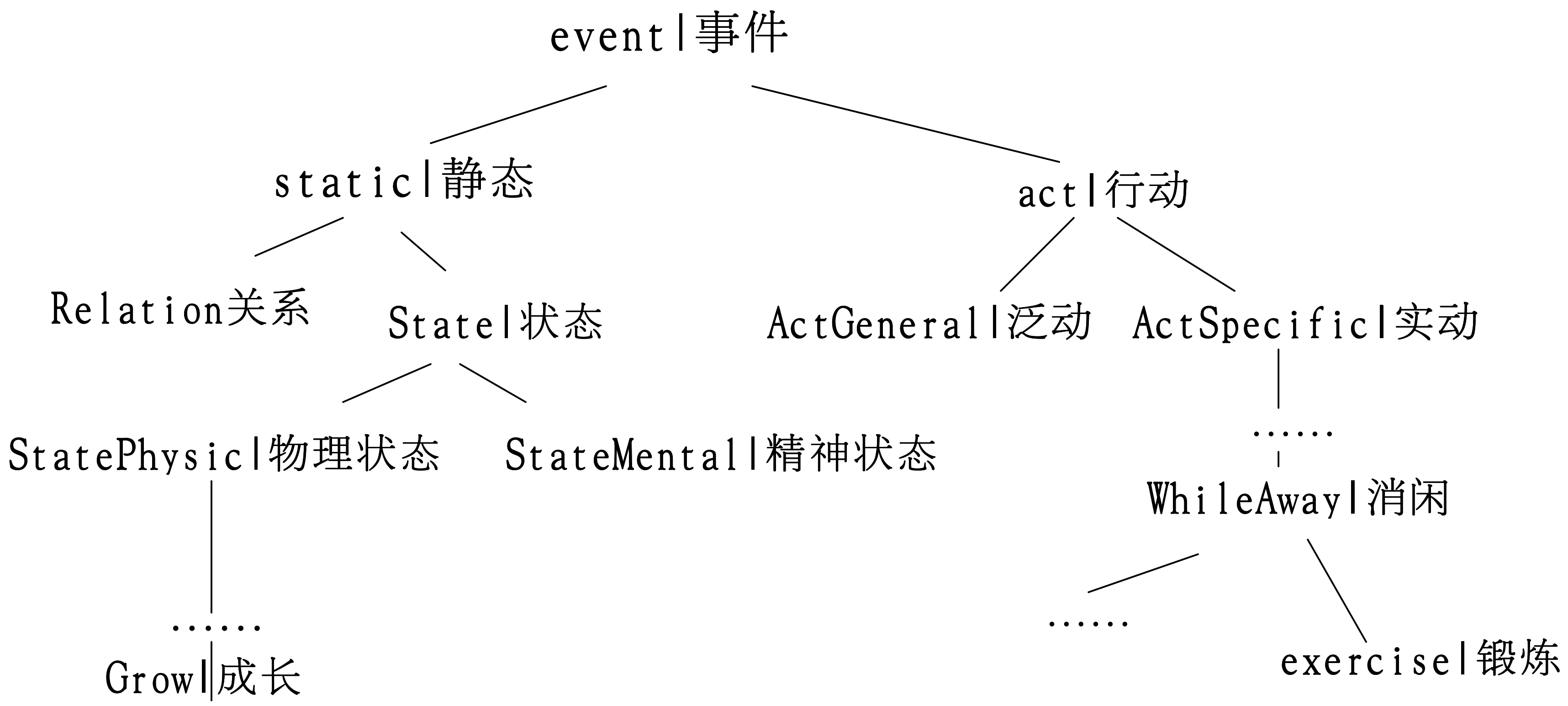
Our method...

- It is a hybrid method, of Knowledge Based and example based, on semantic optimization.
- It takes the *Hownet as* the resource of semantic knowledge including:
  - Taxonomy Tree
  - semantic collocation examples
- It gets hint from the Example-besed method.
- The purpose of it is to solve lexical and structural problems.

Introduction to *Hownet…*

- Hownet is a machine dictionary built as a network for language processing. It is contributed by Professor Dongzheng Dong
- It is actually A large knowledge base using
  - about 1500 primary concepts and their relationships to describe word meaning
  - taxonomy trees to describe the relationship between concepts.
  - collocation examples to discriminate one word meaning from others.

Taxonomy Tree in the *Hownet*



Event taxonomy tree

Lexicon representation in Hownet

- The lexicon representation of a word item:
  - NO.= Serial number of words or phrases
  - [W_X= surface form of words or phrases
  - G_X= Syntactic category of words or phrases
  - E_X= examples of words or phrases]+
  - DEF= concept definition

*__Here__ X in W_X, G_x and E_x will be instanciated as C and E to represent Chinese and English respectively.

A Entry of a word item in Hownet

"打" as "打球"

- NO.＝ 017140
- W_C＝ 打
- G_C＝ V
- E_C＝ ～网球，～牌，～秋千，～太极，球～得很棒
- W_E＝ play
- G_E＝ V
- E_E＝
- DEF＝ exercise|锻炼，sport|体育

The Main Idea of The method

- Find a corresponding collocation word set for every notional word in a dependent tree of the syntactic structure
- calculate the similarity degree of the word set and the example set of each semantic entry of the notional word
- take the entry that gets the highest similar degree as the correct explanation of the notional word in the current structure.
- The score of a dependent tree is the sum of all of its word nodes which are explained by the meaning entry with highest similar degree.
- With ambiguity structures, choose the one that gets the highest score as the final result.

The Main steps of The method

- Find the context-windows of each notional word in the structure.
- calculate the similarity degree of the notional word set and the example set of each semantic entry of the notional word:
  - Calculate semantic similarity between semantic primatives ==> using the taxonomy tree
  - Calculate semantic similarity between words ==> using the DEF of a entry
  - Calculate semantic similarity between examples of the word meaning entry and the words' context window ==> using the example of a entry.

Example of the disambiguation algorithm

- Input phrase:维修图书馆的空调

Syntax parsing results：

  - R1:（（维修/v 图书馆/n）的 空调/n）/np
  - R2:（维修/v （图书馆/n 的 空调/n））/vp

	维修	图书馆	空调	TOTAL
R1	0.45	39.88	23.41	63.74
R2	88.46	0	14.39	102.85

finally selected result is : R2

Experimental result

- We did a experiment with small sentence corpus:The Number of Sentences in the Corpus: 800 sentences.It covers the sentences in the Chinese text book of first and second year of primary school.
- The number of sentences that can get the correct parsing result after disambiguation/ the total number of sentences in the corpus CR=91.4%
- We are going to do more experiment using corpus with large scale.

Experimental result

- Table. 2 the average evaluating time under different length and entry number (Unit time）

	1-5	6-9	10-13	14-17	18-21
1-3	17	32	44	56	84
4-6	70	101	163	217	283
7-9	200	362	562	767	1001

6. About our partners

- Our partner, the MT Group in the University of Science and Technology of National Defence, lead by Academism Chen Huowang and Dr. Shi Xiaodong and Wangting, is one of the pioneers of MT in China. Funded by 863 Foundation, They have had several remarkable achievements:
  - MATRIX English-Chinese MT system
  - ICENT Chinese-English MT system
  - And Neon, the bi-directional English-Chinese MT system.
- Some features of these system, especially the quality and the speed of their English-Chinese MT system is in the leading position in China and even worldwide.

About our partners

- They built a internet explorer with the functionality of English-Chinese and Chinese-English translation, and the first www station for MT in China.
- Their contributions are noted in the leading position of Chinese contributors in the Software Cycropedia for Natural Language Processing and Machine Translation by a English scholar of MT, John Hutchins.

7. The directions and plans of our future research and development

In order to make further improvement in the quality and speed of machine translation:

- Study the practical knowledge representation, organization and utilization scheme, including the Interlingua.
- Better combine the rule-based and corpus-based method
- Inprove the reasoning mechanism to analyze a sentence in a discourse context.

The directions and plans of our future research and development

At the same time, we will try to build several application systems:

- Better Chinese-English MT system
- UNL encoder.
- Machine added translation system