Presentation on the 11th China-Japan
Natural Language Processing Workshop

Multilingual Machine Translation Research in CAS-ICT

LIU Qun （刘群）　　LÜ Yajuan （吕雅娟）

Miyazaki, Japan, 2011-10-30

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

Outline

- Background
- Challenges
- Progress
- Conclusion

Minority Languages in China

- China is a big country with 56 nationalities.
- Besides the Chinese language, there are more 80 minority spoken languages, and about 30 minority written languages in China.
- Some of the minority written languages have a very long history, such as Tibetan, Mongolian, Yi language, while some are new created written languages.

Minority Languages in China

- Some minority languages are the official languages in the autonomous regions.
- The minority languages with population more than one million:

Mongolian
蒙古语

Uygur
维吾尔语

Tibetan
藏语

Kazak
哈萨克语

Korean
朝鲜语


南亚语系
南岛语系
印欧语系
壮侗语族
朝鲜语
满－通古斯语族
苗瑶语族
蒙古语族
藏缅语族
突厥语族
0 250 500
公里
中国语言图
汉语方言
1 官话
2 闽语
3 赣语
4 客家
5 徽语
6 晋语
7 平话
8 吴语
9 湘语
10 粤语
多话
韶州土话
儋州话
中国
苏联
蒙古
朝鲜
日本
菲律宾
印度
尼泊尔
缅甸
老挝
越南
泰国
台湾
内蒙古自治区
新疆维吾尔自治区
西藏自治区
青海
甘肃
宁夏
陕西
山西
河北
辽宁
吉林
黑龙江
山东
河南
湖北
湖南
四川
重庆
云南
贵州
广西壮族自治区
广东
福建
江西
安徽
江苏
浙江
上海
海南
北京
天津
内蒙古自治区
新疆维吾尔自治区
西藏自治区
广西壮族自治区
宁夏回族自治区

Foreign Languages in China

- Besides English, the market demand of translation between Chinese and other foreign languages also increase very rapidly
- The demands in the translation market include:

English

Japanese / Korean

German / French /Russian

Swedish / Dutch / Spain / Italian / Arabic

Thai / Malay / Vietnamese / India

Multilingual Machine Translation Research in China

- In China, most machine translation research focus on between Chinese and English
- Only few research groups did research or development on machine translation between Chinese and Japanese or other languages
- We think it is time for the us to pay more attention to minority languages and other foreign languages besides English

Outline

Background

Challenges

Progress

Conclusion

Challenges in Multilingual MT

- Resource Scarcity
- Multiple Encoding and writing systems
- Spelling Check
- Sentence and Word Segmentation
- Morphological Analysis and Generation
- Syntactic Parsing
- Translation Modeling

Resource

- Bilingual Corpus
  - Collecting Existing Data: Limited Data
    - Purchasing
    - Web Crawling
      - Domain Limit
      - Source Bias
  - Translating: high cost

Resource

- Translation Dictionary
  - General
  - Proper Noun
    - Person names
    - Location names
    - ...

Resource

- Tools
  - Article Alignment
  - Sentence Alignment
  - Word Alignment

Language Independent

Encoding and Writing System

- Some minority languages have many different encoding system or writing system.
- Encoding system:
  - Unicode
  - Many language has its own encoding system
- Writing system:
  - Traditional system
  - Romanization system
- The existing corpus usually use different encoding system.
- It is very important for the detection the encoding system and writing system and transfer between these systems.

Writing System for Mongolian

- For Mongolian, there are two different writing systems
  - Traditional writing system:
    - Traditional alphabets
    - Direction: Top Down, Left to Right
    - Being used in China
  - Cyrillic writing system
    - Cyrillic alphabets (like Russian)
    - Direction: Left to Right, Top Down
    - Being used in Mongolia

Encoding System for Traditional Mongolian

- For traditional Mongolian, there are many different encoding systems.
- Problem:
  - A character may have different glyph (shape) when it appear in different position in a word
  - Different character my have the same glyph (shape)
- There are two different kinds of encoding systems:
  - Glyph-based Encoding system
  - Pronunciation-based Encoding System

Encoding System for Traditional Mongolian

- Glyph-based Encoding system
  - One glyph, one code
  - Exact description of the shape
  - Linguistically incorrect
  - Broad used in electronic publication system
- Pronunciation-based Encoding System
  - One alphabet, one code
  - Linguistically correct
  - Adopt by Unicode and Romanized writing systems

Encoding System for Traditional Mongolian

- In machine translation, we will adopt the pronunciation-based encoding systems, since if we use the glyph-based encoding system, it will be impossible for us to do morphological analysis
- Mongolian is agglutinative language, where each word may have more 2000 different variations.
- Without morphological, there will be too many OOVs, which will bring great difficulties in MT

Encoding System for Traditional Mongolian

- It is a very difficult problem to transform between the glyph-based encoding systems and the pronunciation-based encoding system.
- Such a transform problem is a many-to-many mapping problem

Spelling Check

- Spelling check is a very big problem in some languages, such as Uygur, and Cyrillic Mongolian.
- Even for the famous writers, there are also many spelling mistakes in there works.
- For human readers, such kind of spelling mistake will make no uncomfortable.
- However, for a computer, spelling mistake is a great challenge since it will make a word unrecognized or wrongly recognized.
- So we will have to do spelling check for these kinds of languages for real application.

Sentence Segmentation

- For some languages, such as Tibetan and Thai, there are no explicit boundary between sentences.
- So sentence segmentation is a big problem for the beginning of many NLP applications in these languages.

Word Segmentation

- Word segmentation are also problems for the languages such as Tibetan and Thai.
- The technologies for Chinese word segmentation can also be used for these languages.
- However, there are also some difference.
- For example, Thai is a alphabet-based language rather than a character-based language like Chinese, so the character-labeling-based word segmentation approach can not be used directly

  character-labeling ==> alphabet-labeling

Morphological Analysis

- For morphologically rich languages, such as Mongolian, Uygur, Japanese and Korean, morphological analysis is very important.
- For Mongolian and Uygur, there are no existing morphological analysis software.
- To develop a morphological system for a new language, we need help from the linguistics for the specific language, rather than a normal native speaker of that language, or we need a large stemming corpus.

Morphological Generation

- Morphological generation is also very important for machine translation which takes a morphologically rich language as target language.
- Very few researchers pay attention on this issue currently.

Syntactic Parsing

- Syntactic parsing will benefit machine translation between the language pairs with big differences.
- For the minority languages, there are almost no researches on parsing technology, and almost no treebank is available.
- For the morphologically rich languages, we wonder if the popular PCFG-like approach will get satisfied result. The problem is how to use the rich morphological information.

Translation Modeling

- Current statistical translation model regards each word as a symbol.
- It is not nature to use the rich morphological information for such languages.

Challenges in Multilingual MT

	Resou.	Spelli. Check	Sent. Segm.	Word Segm.	Encod.	Morph. Analy.	Morph. Gener.	Synt. Parsing	Trans. Modeli.
Mongolian	★	★			★	★	★	★	★
Tibetan	★		★	★	★			★	
Uygur	★	★			★	★	★	★	★
Korean	★			★		★	★	★	★
Vietnamese	★	★						★	
Thai	★		★	★				★	

Outline

Background

Challenges

Progress

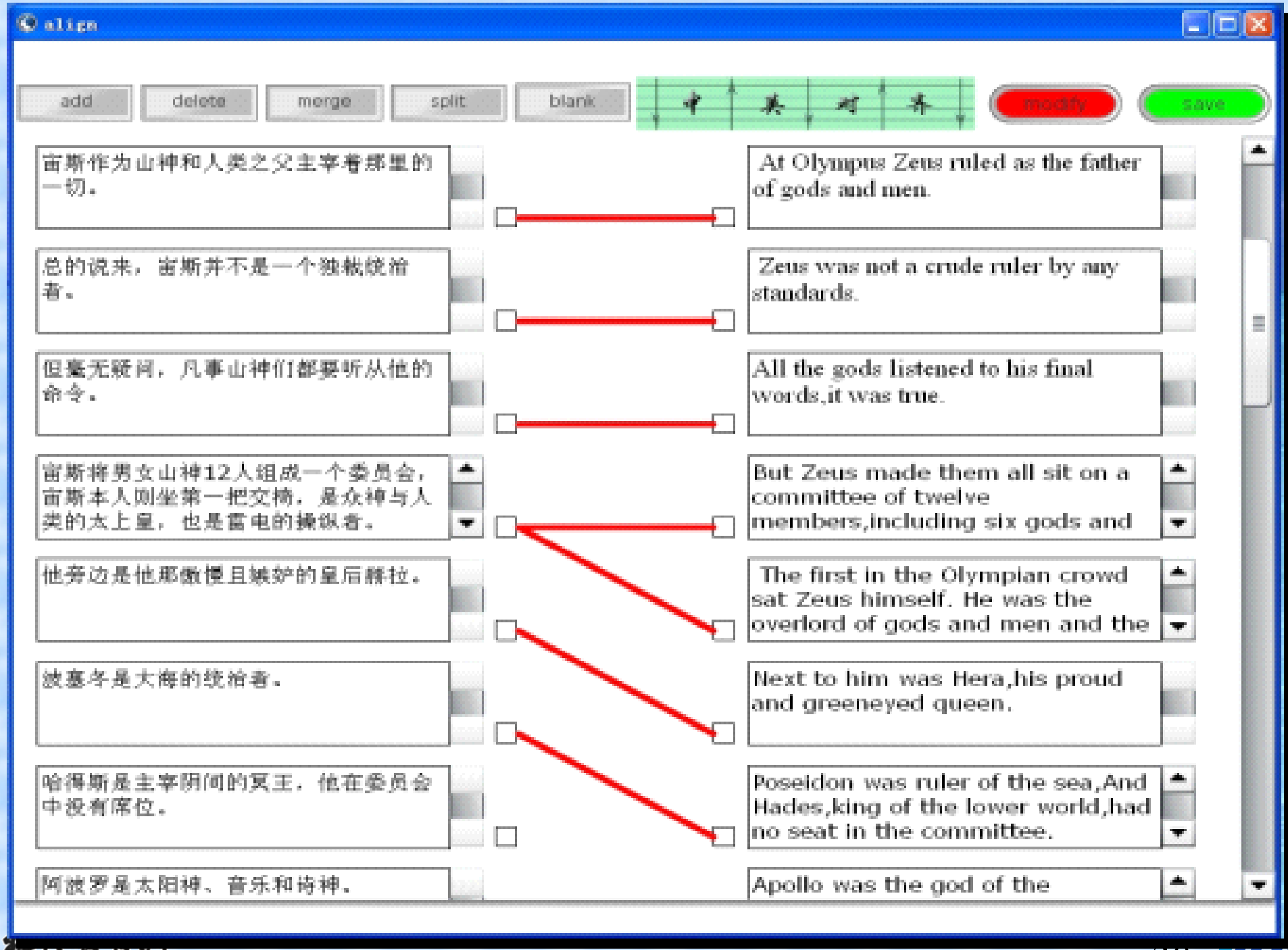
Conclusion

Multilingual MT Research in ICT

- Minority Language MT
  - Mongolian ---- Inner Mongolian University
  - Tibetan ---- Qinghai Normal University
  - Uygur ----Xinjiang University
- Other Language MT
  - Korean
  - Russian
  - Thai
  - Vietnamese
  - Japanese

Resources

- Bilingual Corpus & Translation Dictionary
  - Uygur-Chinese
  - Mongolian-Chinese
  - Tibetan-Chinese
  - Russian-Chinese
  - Thai-Chinese
  - Vietnamese-Chinese
- Annotation Tools


align
add
delete
merge
split
blank
modify
save
宙斯作为山神和人类之父主宰着那里的一切。
At Olympus Zeus ruled as the father of gods and men.
总的说来，宙斯并不是一个独裁统治者。
Zeus was not a crude ruler by any standards.
但毫无疑问，凡事山神们都要听从他的命令。
All the gods listened to his final words,it was true.
宙斯将男女山神12人组成一个委员会，宙斯本人则坐第一把交椅，是众神与人类的太上皇，也是雷电的操纵者。
But Zeus made them all sit on a committee of twelve members,including six gods and
他旁边是他那傲慢且嫉妒的皇后赫拉。
The first in the Olympian crowd sat Zeus himself. He was the overlord of gods and men and the
波塞冬是大海的统治者。
Next to him was Hera,his proud and greeneyed queen.
哈得斯是主宰阴间的冥王，他在委员会中没有席位。
Poseidon was ruler of the sea,And Hades,king of the lower world,had no seat in the committee.
阿波罗是太阳神、音乐和诗神。
Apollo was the god of the

Encoding System & Writing System

- Mongolian
  - Romanization <=> Unicode (Traditional Mongolian)
  - Traditional Mongolian <=> Cyrillic Mongolian (in development)
- Uygur
  - Traditional Uygur <=> Romanization Uygur
- Tibetan
  - Banzhida code <=> Unicode
  - Other 7 Encoding system <=> Unicode

Spelling Check

- Uygur Spelling Check System

Mongolian Morphological Analysis

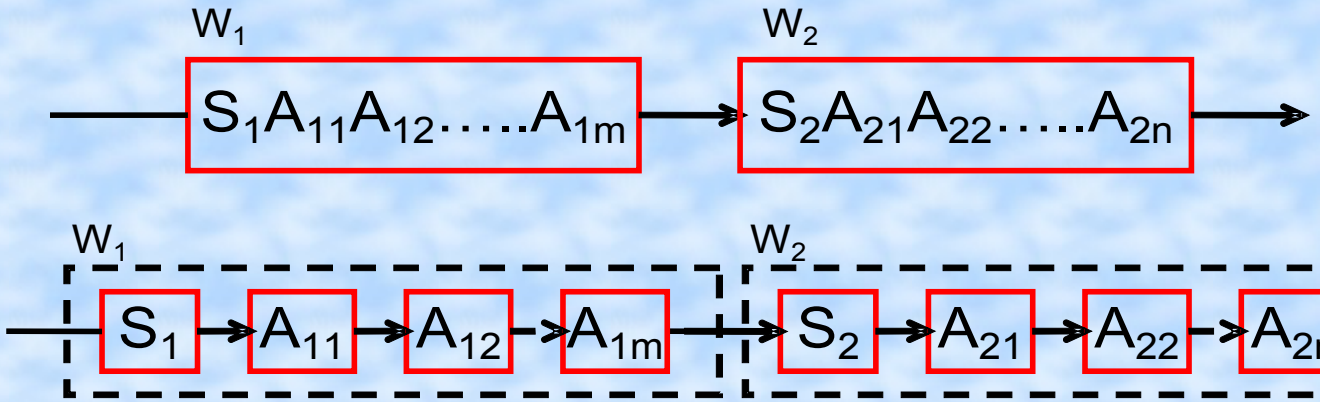
- A sample: HUURNILDU/HU-DU

HUURNI/Ve2 + LDU/Fe3 + HU/Ft12 - DU/Fc21

Stem

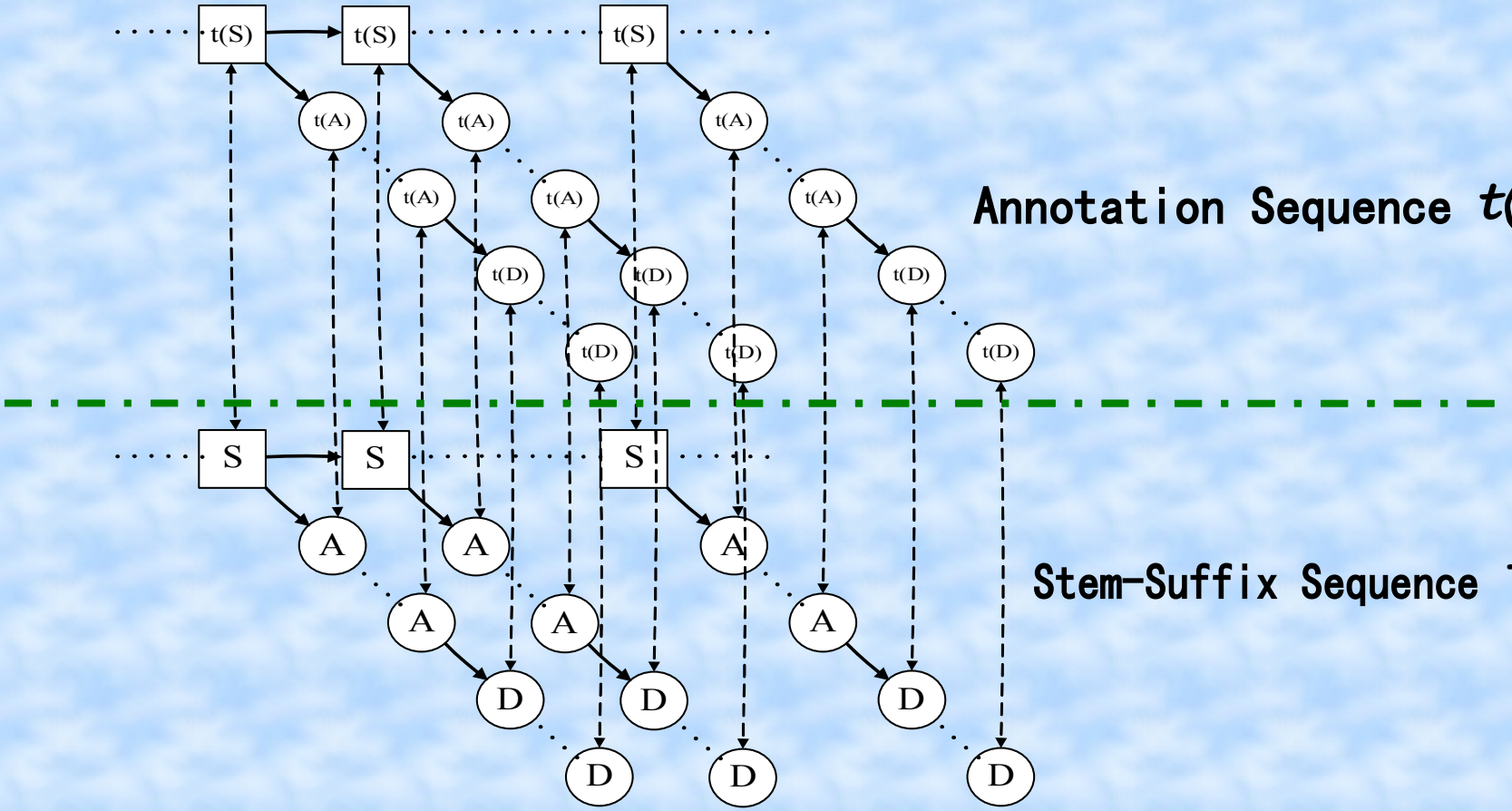
Attached Suffix

Unattached Suffix



Mongolian Morphological Analysis

- A Directed-Graph-based Model for Joint Stemming and tagging



$$P(G) = P(T) \times P(t(T)) \times P(T, t(T))$$

Mongolian Morphological Analysis

- Experiment Result
  - Corpus: 200K Words: Stemming + Annotation
  - Training: 90%, Development: 5% Test: 5%
  - Performance:
    - Stemming: 98%
    - Stemming+Annotation: 96%

设置	Pw+t	Pw-t	Fsa+t	Fsa-t
原有方法	93.0	95.1	93.0	94.7
+判别式词干词缀切分	95.2	97.8	96.1	98.0

Uygur Morphological Analysis

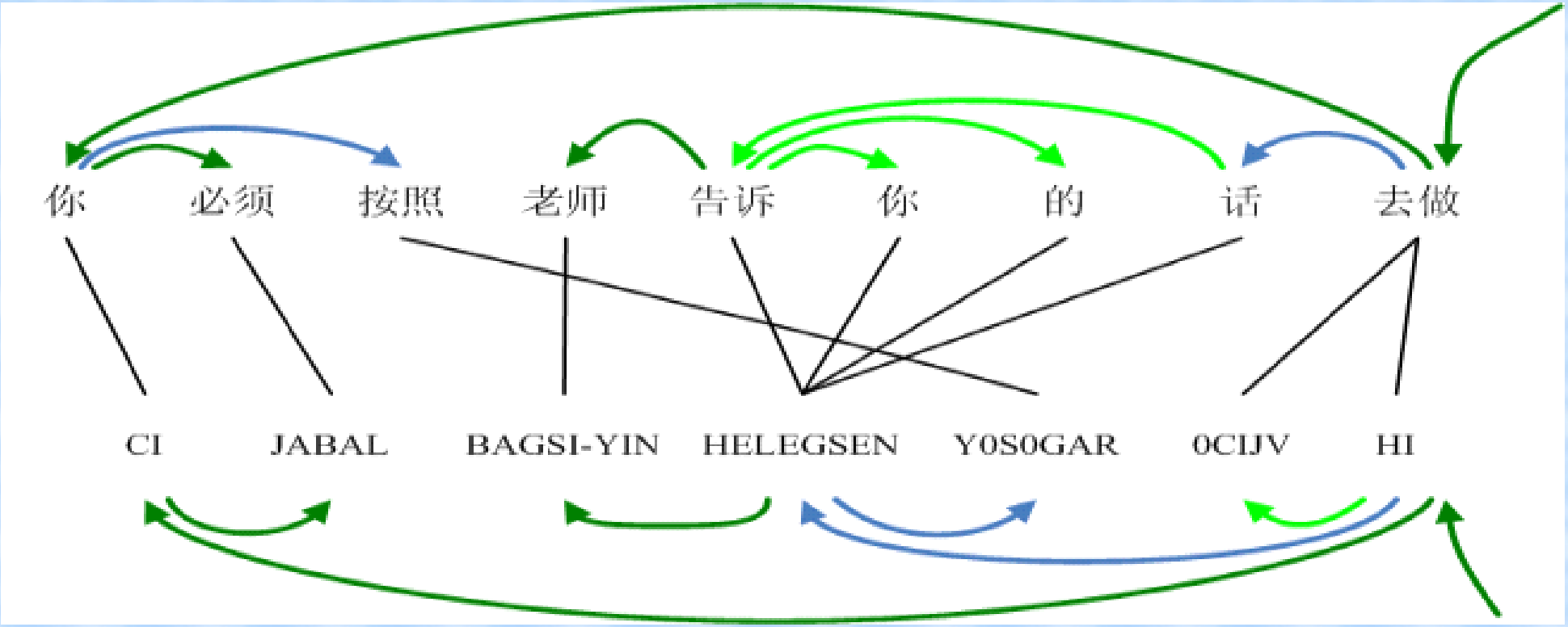
- Experiment Results
  - Corpus: 1M Words: Stemming + Annotation
  - Training: 90%, Development: 5% Test: 5%
  - Performance:
    - Stemming: 94.7%
    - Stemming+Annotation: 92.6%
- Chinese Person Name Recognition in Uygur Text
  - Chinese Person Names may have inflection (in development)

Tibetan Sentence Segmentation and Word Segmentation

- Tibetan Sentence Segmentation: 97%
- Tibetan Word Segmentation: 96%

Mongolian Dependency Parsing

- Mongolian Dependency Parsing based on Bilingual Restriction



Multigrain Machine Translation

- Uygur → Chinese

Grain	BLEU
Word	0.3962
Stem	0.4024
Morpheme	0.3989
Multigrain	0.4165

- Mongolian → Chinese

Grain	BLEU
Word	0.1373
Stem	0.1610
Morpheme	0.1615
Multigrain	0.1818

Outline

Background

Challenges

Progress

Conclusion

Conclusion

- It's time for us to pay more attention to languages other than English
- Many new challenges will be encountered when we work on multilingual machine translation
- We have made some progress in multilingual machine translation
- However, there are still a long way to go ...

谢谢！

Thank you!

ありがとう

Welcome to our website：

http://nlp.ict.ac.cn/new/

http://nlp.ict.ac.cn/new/demo/demo-MT-index.php