National Air and Space Intelligence Center

Engine-specific Phrase Level User Parallel Corpora

Mr. Weimin Jiang
NASIC/GXKT
28 October, 2012

This Briefing is:

UNCLASSIFIED

Introduction

Significance of parallel corpora

- Playing an ever-increasing role in statistical translation (Koehn, 2005)
- Invaluable resources for many NLP applications, such as machine translation, multilingual lexicography, and cross-lingual information retrieval, (Lu et al, 2009)
- Critical resources for extracting translation knowledge in machine translation (Zhang et al).
- Wide applications in Chinese-English information processing, bilingual lexicography, language research and teaching (Chang, 2002).

Introduction

Focus of parallel corpora development

- Automated construction
- Wide coverage

"Automatic and semi-automatic techniques for lexical acquisition are more critical now than ever before as it becomes infeasible to produce adequate semantic representations on a large scale by human labor alone." (Bonnie Jean Dorr et al, 2001)

Introduction

Why manual creation of parallel corpora

- Accuracy
- Fluency

This presentation offers some strategies that might be used in manual development of engine-specific phrase level Chinese-English parallel corpora, hoping to trigger further research and discussion.

Introduction

Why manual creation of parallel corpora

- Lack of inflections in Chinese language
- Lack of rigid grammar in Chinese language
- Very different positions of prepositional and adjective phrases
- Very different sentence structure
- Implicit meaning in both Chinese and English
- Multiple meanings

Introduction

Why engine-specific

Source:
他引用(quote)新华社报道(report)称(say),中国此次"陆基中段反导拦截技术"试验是在"中国境内"进行的。试验达到了预期目标。

HT:
He quoted a Xinhua news report saying that this test of Chinese "ground-based midcourse missile interception technology" was conducted "within Chinese territory". The test has achieved its anticipated goal .

Introduction

Why engine-specific

Systran:

He quotes Xinhua News Agency to report said that China this time "ground-based midcourse missile interception technology" experiment is conducted in "China". The experiment has achieved the anticipated target .

Google:

He quoted the Xinhua News Agency reported that the "ground-based midcourse missile interception technology" test in China conducted. Test has achieved the expected goals.

Microsoft:

He cited the Xinhua News Agency reported that China's "ground-based Midcourse missile interception technology" test was "Republic of China". Test and achieved the desired objective.

Introduction

Why engine-specific

- **Parsing**
- **Programming**
- **Phrase and sentence organization**

Due to the difference in MT engine mechanism, we might need first to identify incorrect MT translation and then create engine-specific phrase-level parallel corpora of frequently repeated phrases for each domain.

Introduction

Why engine-specific

Systran:

He quotes Xinhua News Agency to report said that China this time "ground-based midcourse missile interception technology" experiment is conducted in "China". The experiment has achieved the anticipated target.

MT with UD

He quotes a Xinhua report saying, this Chinese" ground-based midcourse missile interception technology" test was conducted within Chinese territory. The test has achieved the anticipated goal.

Strategies

Strategies for engine-specific parallel corpora

- **Associating words**
- **Defining verbs**
- **Clarifying meaning or simplifying structure**

Strategies

Associating words

从信息流动的角度, 现有的许多搜索引擎技术都是不可靠的。

MT:

From information flow's angle, existing many search engine technologies are not reliable.

从信息流动的角度讲(speak), 现有的许多搜索引擎技术都是不可靠的。

HT:

(Speaking) from the angle of information flow, many existing search engine technologies are not reliable.

MT without UD:

From the angle of information flow said that existing many search engine technologies were unreliable.

MT with UD:

From information flow's angle, many existing search engine technology is unreliable. (UD entries: 角度讲=angle, 现有的许多=many existing)

Strategies

Associating words

从信息流动的角度讲**(jiang/speak)**, 现有的许多搜索引擎技术都是不可靠的。

从信息流动的角度**说(shuo/say)**, 现有的许多搜索引擎技术都是不可靠的。
从信息流动的角度**谈(tan/talk)**, 现有的许多搜索引擎技术都是不可靠的。

UD entries: **角度** *讲*/**说**/**谈**=*angle*
现有的*许多*/**一些**=many existing/some existing

Association and verification

驰名世界的, 闻名世界的, 世界驰名的,举世闻名的, 享誉世界的,世界闻名的, etc. = world-famous, but only the first two are not correctly translated.

Strategies

Associating words

- Using synonyms of the same part of speech (e.g. **角度**讲 and **角度**说)
- Changing word order, (e.g. many existing: 现有的许多 and 许多现有的)
- Eliminating or adding particles (e.g. operation procedure: **操作程序要求** and **操作程序的要求**)
- Changing meaning to the opposite (e.g. rockets with carrying capacity ***over*** 20 tons/20**吨以上运**载能力的火箭 and rockets with carrying capacity ***under*** 20 tons/20**吨以下运**载能力的火箭)
- Adding or removing inessential words (e.g. the most ~~most~~ rigid **最最**严格的/**最**严格的)

Strategies

Defining verbs

中国研制(develop)的高速铁路试验车将于明年进行(have/conduct)速度试验。

HT:
The high-speed test train developed by China will see a speed test next year.

Without UD:
China develops the high-speed railroad testing car will carry on the speed trial next year.

With UD:
The Chinese developed high-speed railroad testing car will carry on the speed trial next year. (with UD entry: 研制的=developed)

Strategies

Defining verbs by expansion

正在研制(develop)的高速列车时速可达(reach)400公里。(passive relation with its logic subject)

HT:
The train being developed can reach a speed of 400 km/h.

MT without UD:
The high-speed train speed that developed may amount to 400 kilometers.

MT with UD:
The high-speed train being developed may amount to 400 kilometers.
(UD entries: 正在研制的高速列车= the high-speed train being developed)

Strategies

Clarifying meaning or simplifying structure

时速**达**(reach)400**公里的高速列**车正在研制(develop)**中**。

HT:
A high-speed train of 400 km/h is being developed.

MT without UD:
The speed reaches 400 kilometers high-speed train to develop.

MT with UD:
The high-speed train of 400km/h is being developed.
(UD entry: 时速达400公里=400km/h)

Strategies

Clarifying meaning or simplifying structure

他在接受(accept)本刊记者采访(interview)时介绍(introduce)说(say), 中国发射(launch)导弹是(is)防御性的。

HT:
In an interview with our reporter, he said China's missile launch is defensive in nature.

MT without UD:
He when accepting this publication reporter interviewed says, China fired the missile is defensive.

MT with UD:
He said during an interview with our reporter, China's missile launch is defensive.
(UD entries: 在接受本刊记者采访时介绍说=said during an interview with our reporter, **中国**发射导弹=China's missile launch)

Compromise

User parallel corpora or user dictionaries can address considerable amount of translation errors, improving fluency and accuracy. But in many cases they can only offer compromises.

Compromise

雷达通常由发射机、发射天线、接收机、接收天线以及显示器组成。

HT:
Radar is generally composed of a transmitter, transmitting antenna, receiver, receiving antenna, and display screen.

MT:
The radar usually is composed of the transmitter, transmitting antenna, receiver, receiving antenna as well as the monitor.

Compromise

由发射机、发射天线、接收机、接收天线以及显示器组成的雷达在军事行动中起到很重要的作用.

HT:
Radar composed of a transmitter, transmitting antenna, receiver, receiving antenna, and display screen plays a critical role in military operations.

MT:
In the military action plays very vital role from the radar that the transmitter, the transmitting antenna, receiver, the receiving antenna as well as the monitor are composed.

Compromise

A user corpus for 由发射机、发射天线、接收机、接收天线以及显示器组成?

Available options are (5!)=120

Conclusion

Strength

- *Significantly improve MT accuracy and fluency*
- *Save time in MT*

Weakness

- *Development is time consuming*
- *Some corrections are beyond reach*

Acknowledgement

This article is based on Mr. David Barber's concept of phrase-level user parallel corpora for Systran translation engine. His constructive advice, suggestion, and other input are greatly appreciated.

Ms. Jin Yang, Systran computational linguist, has offered a lot of insight into Systran MT mechanism, which has greatly contributed to this paper.

Thank you

One more example

Original
一个来自俄亥俄玉米地的家伙向你们表示衷心感谢

MT without UD
Fellow of Ohio cornfield expressed heartfelt gratitude to you.

MT with UD
A guy from Ohio cornfield wants to express his hearty thanks to you.

Questions?

POC: Mr. Weimin Jiang
COM: (937)-522-6191
Weimin.jiang@wpafb.af.mil