FUJITSU

shaping tomorrow with you

富士通研发中心日汉专利机器翻译方法研究

郑仲光

2012/08/16

Outline

FUJITSU

- FRDC统计机器翻译研究现状
- 利用中间语
  - 构建中日翻译词典构建
  - 构建中日翻译系统
- 下一步工作

基础技术

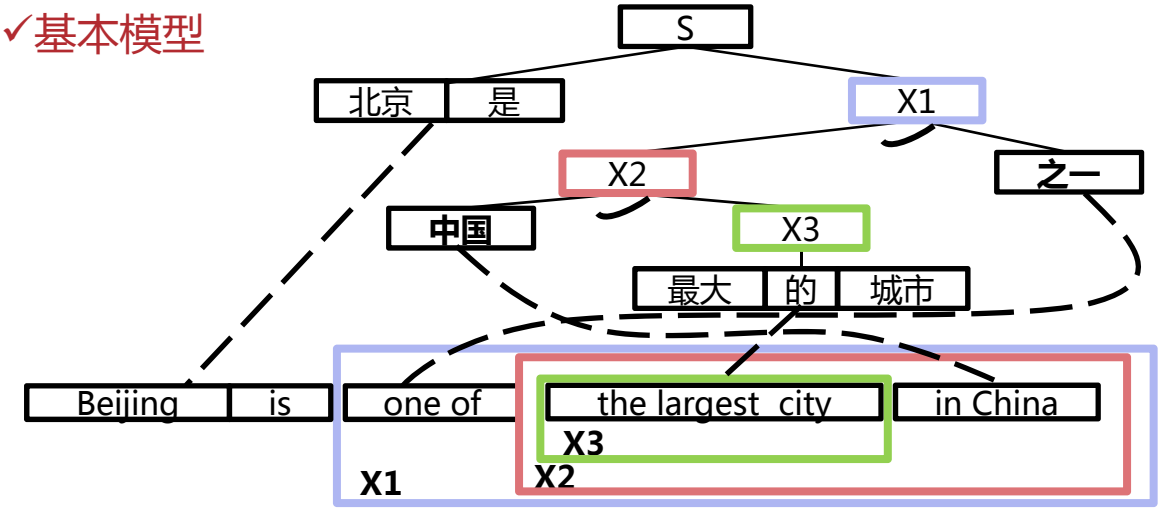
FUJITSU

- Hierarchical Phrase-based SMT(Chiang, 2005)

2009 —

2008 —

✓基本模型



✓研究工作

翻译规则过滤；翻译规则选择；调序规则控制；汉语改写；
多词单元抽取； .....

ACL, COLING, EMNLP, AMTA, PACLIC, MT summit

评测工作及产品

FUJITSU

汉英专利翻译

2011 —

2010 —

2009 —

2008 —

✓评测工作

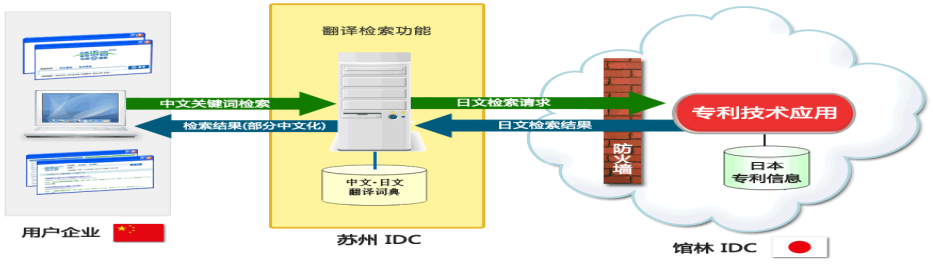
NIST(09)

Chinese—English	
BLEU	8/23

NTCIR(11)

Chinese—English	
BLEU	Human evaluation
8/23	7/23

✓成果：
汉英日三语
专利检索系统

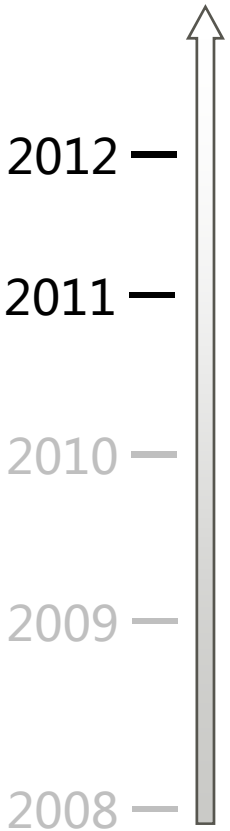


http://www.fpatents.com/patent/

专利翻译

FUJITSU

专利语料资源积累



✓双语资源

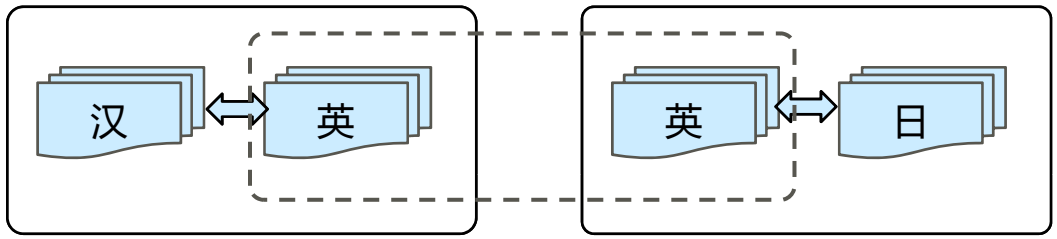
语言	数量	说明
中英	500万	1985-2009年共25年的中英双语专利数据(全部领域)
	100万	NTCIR 评测训练语料
中日	10万	电学专利标题
日英	300万	NTCIR 评测训练语料

现有问题：
日汉语料稀疏，严重影响文本翻译、跨语言检索质量；

解决方案

FUJITSU

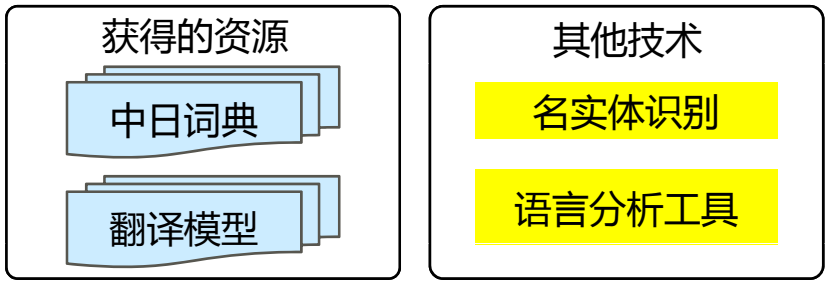
✓语料收集：积极获取
中日语料资源



✓用英文作为中间语

✓目标：

- 获取领域词典：跨语言检索
- 翻译模型：中日文本翻译
- 结合其他NLP技术提高CJ翻译质量



Outline

FUJITSU

- FRDC统计机器翻译研究现状
- 利用中间语
  - 构建中日翻译词典构建
  - 构建中日翻译系统
- 下一步工作

词典的基本单位

FUJITSU

多词单元

- 在特定领域内，惯用词组和专有名词被称为多词单元。 多词单元由两个或两个以上的分词构成。每个多词单元具有统一的词性并且在特定领域内经常被使用。在中英双语对齐语料中，多词单元可以对应英文的一个单词或一个词组。

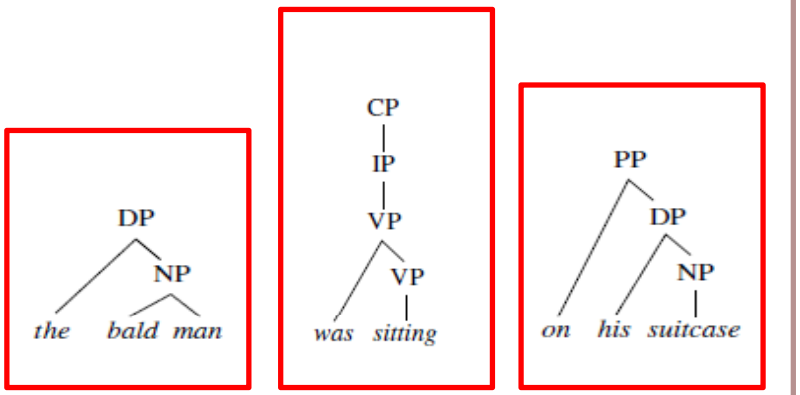
与Phrase, Chunk的区别

- 短语中的介词短语不具有多词表达特点；
- Chunk是对应parsing定的，强调的是在句法上完整而非语义上完整；

Phrase

- face twisted(Absolute Phrase)
- fairly ridiculous(Adjective Phrase)
- a big bright green pleasure machine(Noun Phrase)
- On second thought(Prepositional Phrase)
- With you(Prepositional Phrase)
- may be going away(Verb Phrase)

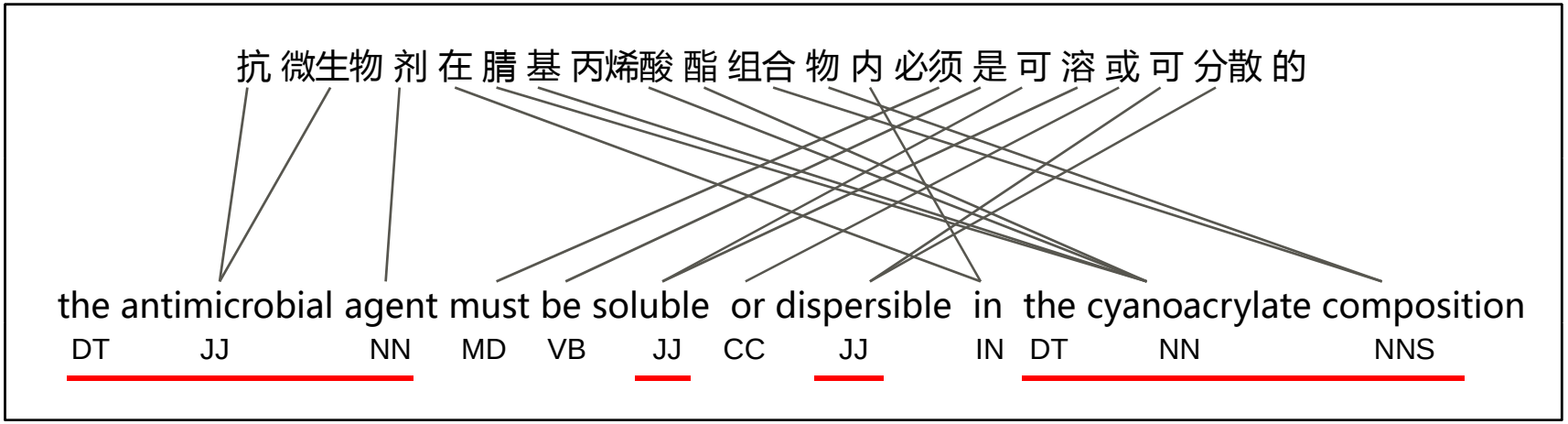
Chunk



基于模板的多词单元抽取

FUJITSU

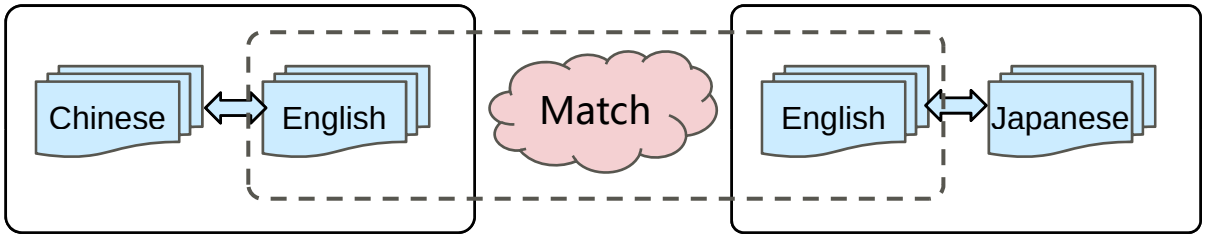
- 词对齐结果



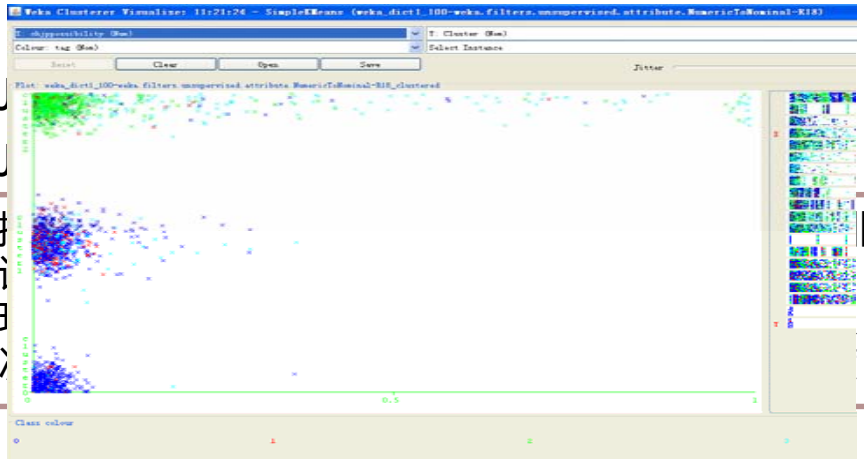
- 模板

In-dex	POS tagging Pattern	In-dex	POS tagging Pattern	Index	POS tagging Pattern
1	(/DT)(/NN)+	4	(/DT)?(/JJ)+(/NN)+	7	((/IN)(/:)(/VV))
2	(/VV)+(/NN)+	5	(/JJ)	8	(/NN)+(/CC)(/NN)+
3	(/NN)(/NN)+	6	(/NN)	10	.etc

利用中间语构建汉日词典



- 多义词
  - -<Chinese>供电<J
  - -<Chinese>当前<J



simpleKMeans算法，选择
组，如此可以过滤掉中间词
中英、日英多词表达对共现
日文多词表达中出现总频次

非多义词）的一

，英文在中文、

- 分词错误
  - -<Chinese>个子 代码<Japanese>サブ コード<English>[sub/NN code/NN]

根据次出现的位置制定停用词表；
“这种”、“与其”、“与该”、“以下”

- 打分模型

利用人工标注的语料训练打分模型，从聚类结果中选取得分为3的词条；

实验

FUJITSU

实验数据&结果

汉英	日英	汉语词数	词条数量	准确率
36万	44万	1	20782	91%
		2	92392	92%

实例

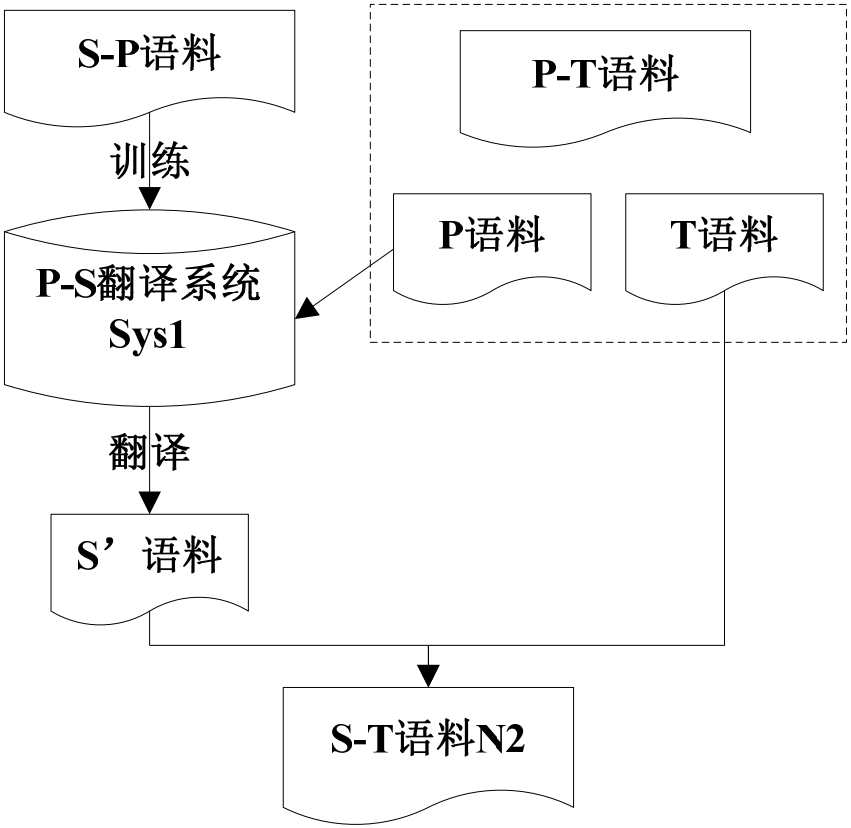
<Chinese>不完善</Chinese><Japanese>不完全</Japanese>
<Chinese>严峻</Chinese><Japanese>厳しい</Japanese>
<Chinese>间歇</Chinese><Japanese>間欠</Japanese>
<Chinese>乙烯</Chinese><Japanese>ビニール</Japanese>
<Chinese>元素铂</Chinese><Japanese>白金</Japanese>
<Chinese>光刻术</Chinese><Japanese>リソグラフィ</Japanese>
<Chinese>光致抗蚀剂</Chinese><Japanese>フォトレジスト</Japanese>
<Chinese>分界符</Chinese><Japanese>区切り</Japanese>
<Chinese>分辩率</Chinese><Japanese>解像度</Japanese>
<Chinese>刚度</Chinese><Japanese>剛性</Japanese>
<Chinese>条形码扫描仪</Chinese><Japanese>バーコード スキャナ</Japanese>
<Chinese>检测信号</Chinese><Japanese>検出 信号</Japanese>
<Chinese>气体混合物</Chinese><Japanese>混合 ガス</Japanese>

Outline

FUJITSU

- FRDC统计机器翻译研究现状
- 利用中间语
  - 构建中日翻译词典构建
  - 构建中日翻译系统
- 下一步工作

基于语料扩展的中间语方法



使用机器翻译的方法，将源语言--中间语语料的中间语端，翻译为目标语言；
使用机器翻译的方法，将中间语—目标语语料的中间语端，翻译为源语言端；
将翻译得到的语料，经过一定的过滤，加入到原来的源语言—目标语言语料中；
使用扩充后的语料，重新训练机器翻译系统，得到系统A

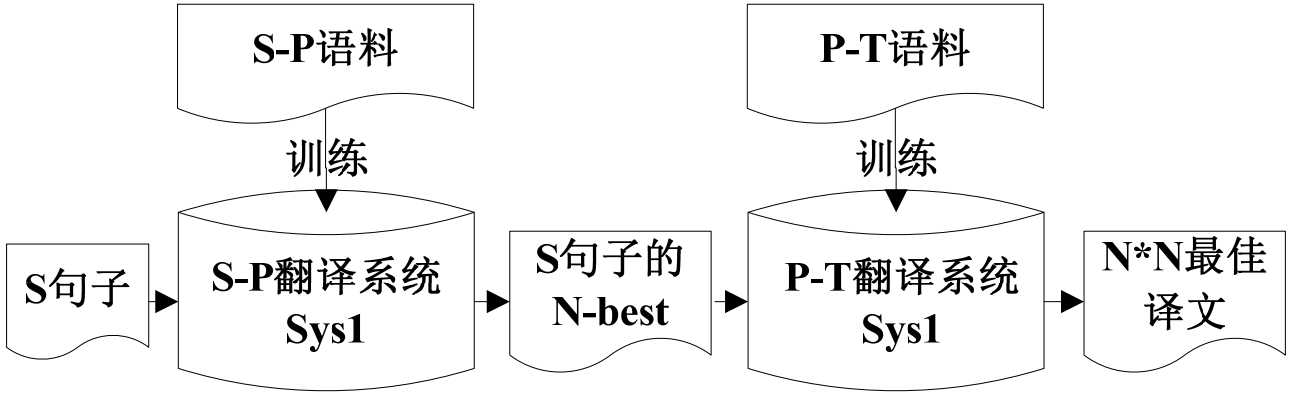
语料过滤，主要参照句子长度比：

$$\alpha < \mathrm{Ls/Lt} < \beta \qquad (1)$$

$$\alpha = \min_{i=1}^{N} \left(\frac{Len(Si)}{Len(Pi)}\right) \qquad (2)$$

$$\beta = \max_{i=1}^{N} \left(\frac{Len(Si)}{Len(Pi)}\right) \qquad (3)$$

基于串联翻译的中间语方法



使用源语言—中间语语料，构建翻译系统Sys1；
使用中间语—目标语语料，构建翻译系统Sys2；
Sys1的输出作为Sys2的输入，两个系统组合成为系统B；
源语言翻译为中间语，取Nbest；
将Nbest用Sys2翻译为N*N个目标语言句子；
最终的译文，为两次翻译中得分之和最高的句子。

$$S(t_{ij}) = \sum_{k=1}^{M} (\lambda_k^{sp} h_k^{sp}(s,p)) + \sum_{k=1}^{N} (\lambda_k^{pt} h_k^{pt}(p,t)) \qquad (4)$$

$$\hat{t} = \arg\max_{t} (S(t_{ij})) \qquad (5)$$

基于规则扩展的中间语方法

$$p(s \mid t) = \sum_{p \in T_{sp} \cap T_{pt}} p(s \mid p) p(p \mid t) \quad (6)$$

$$p(t \mid s) = \sum_{p \in T_{sp} \cap T_{pt}} p(t \mid p) p(p \mid s) \quad (7)$$

$$\phi(s \mid t) = \sum_{p \in T_{sp} \cap T_{pt}} \phi(s \mid p) \phi(p \mid t) \quad (8)$$

$$\phi(t \mid s) = \sum_{p \in T_{sp} \cap T_{pt}} \phi(t \mid p) \phi(p \mid s) \quad (9)$$

在源语言—中间语语料上，抽取翻译规则表T1

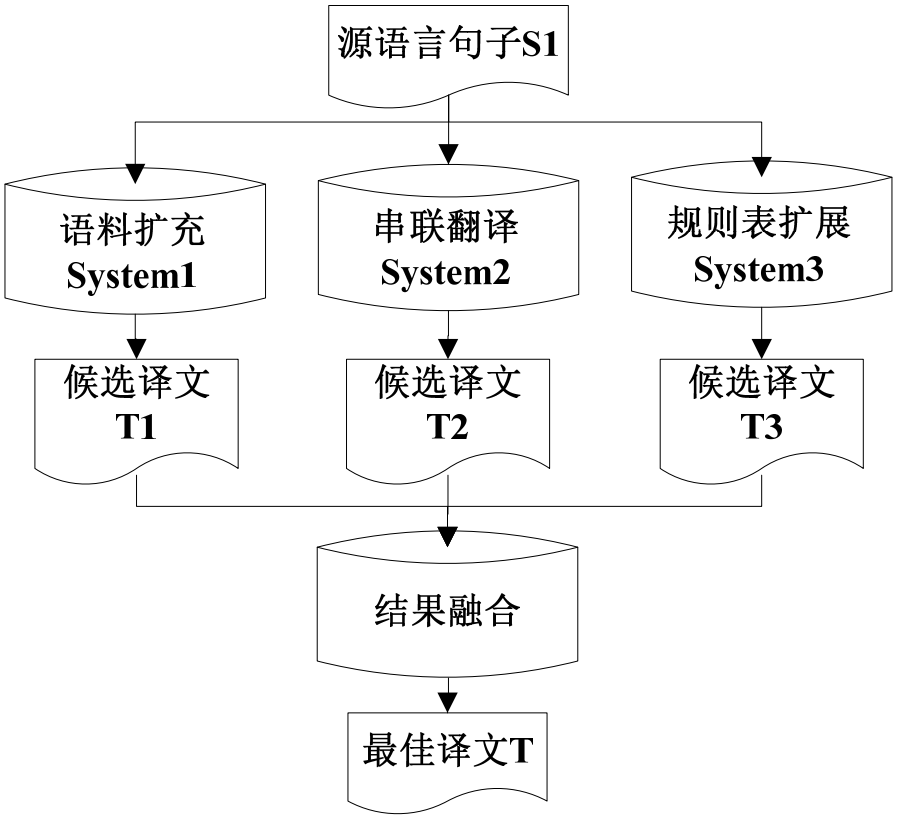
在中间语—目标语语料上，抽取翻译规则表T2

将T1中规则目标端和T2中规则源端相同的规则，组合成新的规则

使用新的规则表，以及其他的翻译资源，构建一个新的机器翻译系统C

为了防止规则表过度膨胀，设置了相同源端所能对应的不同目标端的数量K，K的值可以自己设定（如20）.

结果融合



利用最小贝叶斯风险，将BLEU值作为损失函数，求得最佳翻译

$$E_{mbr} = \arg\min_{E'} \sum_{E} P(E \mid F) L(E, E') \qquad (10)$$

$$E_{mbr} = \arg\min_{E'} \sum_{E} (1 - BLEU(E, E')) \qquad (11)$$

实验

FUJITSU

实验设置

Corpus	Training Set	Dev Set	Test Set
Chinese-English (CE)	6174088	1000	1000
English-Japanese (EJ)	3159152	1000	1000
Chinese-Japanese (CJ)	105615	500	1000

实验结果

系统	BLEU4
Baseline	10.05
语料扩展(+500k)	12.86 (+27.96%)
串联翻译	9.91 (-1.39%)
规则扩展	13.65 (+35.82%)
融合	14.30 (+42.29%)

Outline

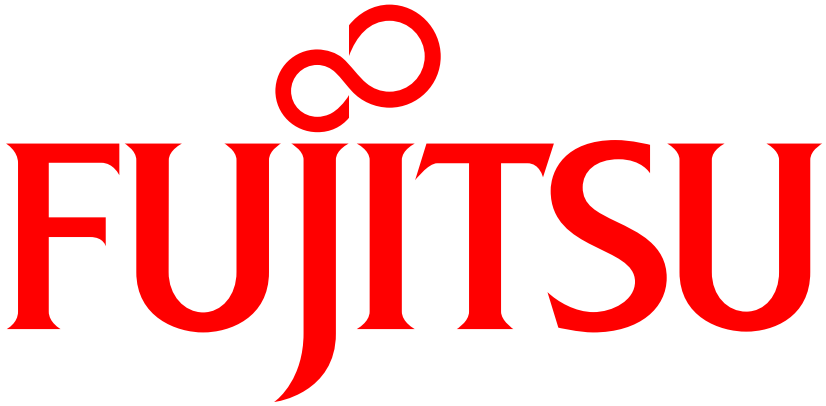
FUJITSU

- FRDC统计机器翻译研究现状
- 利用中间语
  - 构建中日翻译词典构建
  - 构建中日翻译系统
- 下一步工作

未来工作

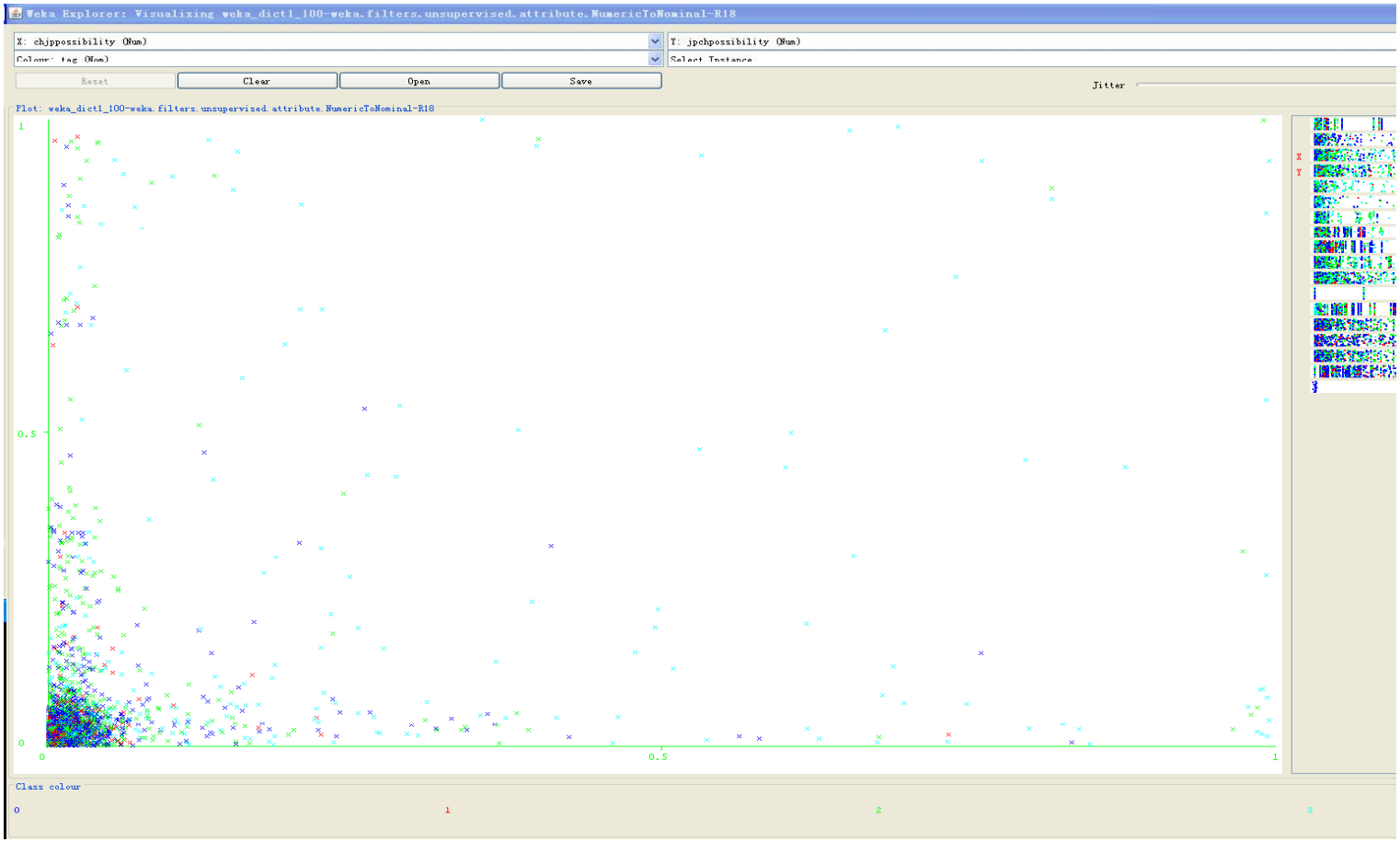
FUJITSU

- 汉日词典构建
  - 提高词条抽取的召回率：
    - 多义词
- 汉日文本翻译
  - 收集语料；
  - 改进方法：
    - 日汉翻译规则的获取；
    - 结合日语分析工具；



shaping tomorrow with you

原始数据



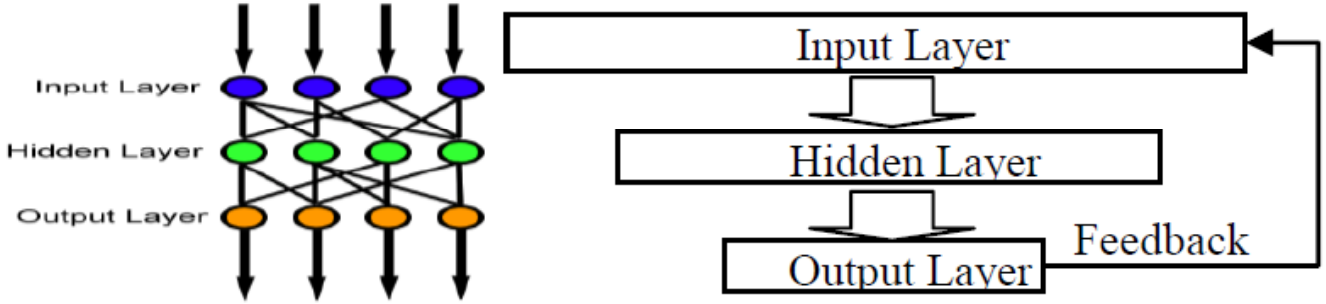
X轴为中日翻译概率，Y轴为聚类结果，根据人为标注着色，2,3为可以接受的词条，分别为绿色和浅蓝色

聚类结果

基于ANN的多词单元抽取

FUJITSU

- 基本模型



- 特征

Chi-nese	Eng-lish(SMT)	Attribute1	Attribute2	Attribute3	…	Attribute154	Labels
最初	initial/JJ	1	2	23	…	False	False
施用	apply-ing/VBG	2	2	6	…	Feedback	False
引	pri-mer/NN	3	1	14	…	Feedback	True
物	pri-mer/NN	4	1	14	…	Feedback	True

总体流程