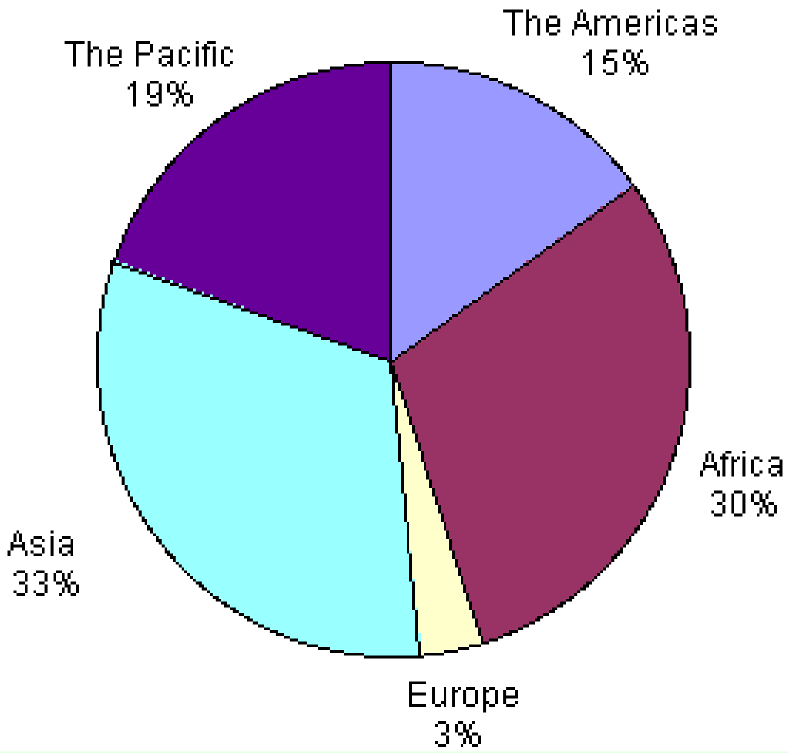
Multilingual Language Processing

Pascale Fung
pascale@ee.ust.hk
Human Language Technology Center
Department of Electronic & Computer Engineering
Hong Kong University of Science and Technology
http://www.ee.ust/~pascale

Most spoken languages in the world

#	Language	Speakers (in millions)	#	Language	Speakers (in millions)
1	Mandarin	1051	11	Japanese	127
2	English	510	12	German	123
3	Hindi	490	13	Farsi/Persian	110
4	Spanish	425	14	Urdu	104
5	Arabic	255	15	Punjabi	103
6	Russian	254	16	Vietnamese	86
7	Portuguese	218	17	Tamil	78
8	Bengali	215	18	Wu Chinese	77
9	Malay	175	19	Javanese	76
10	French	130	20	Turkish	75

Geographic distribution of the world's languages



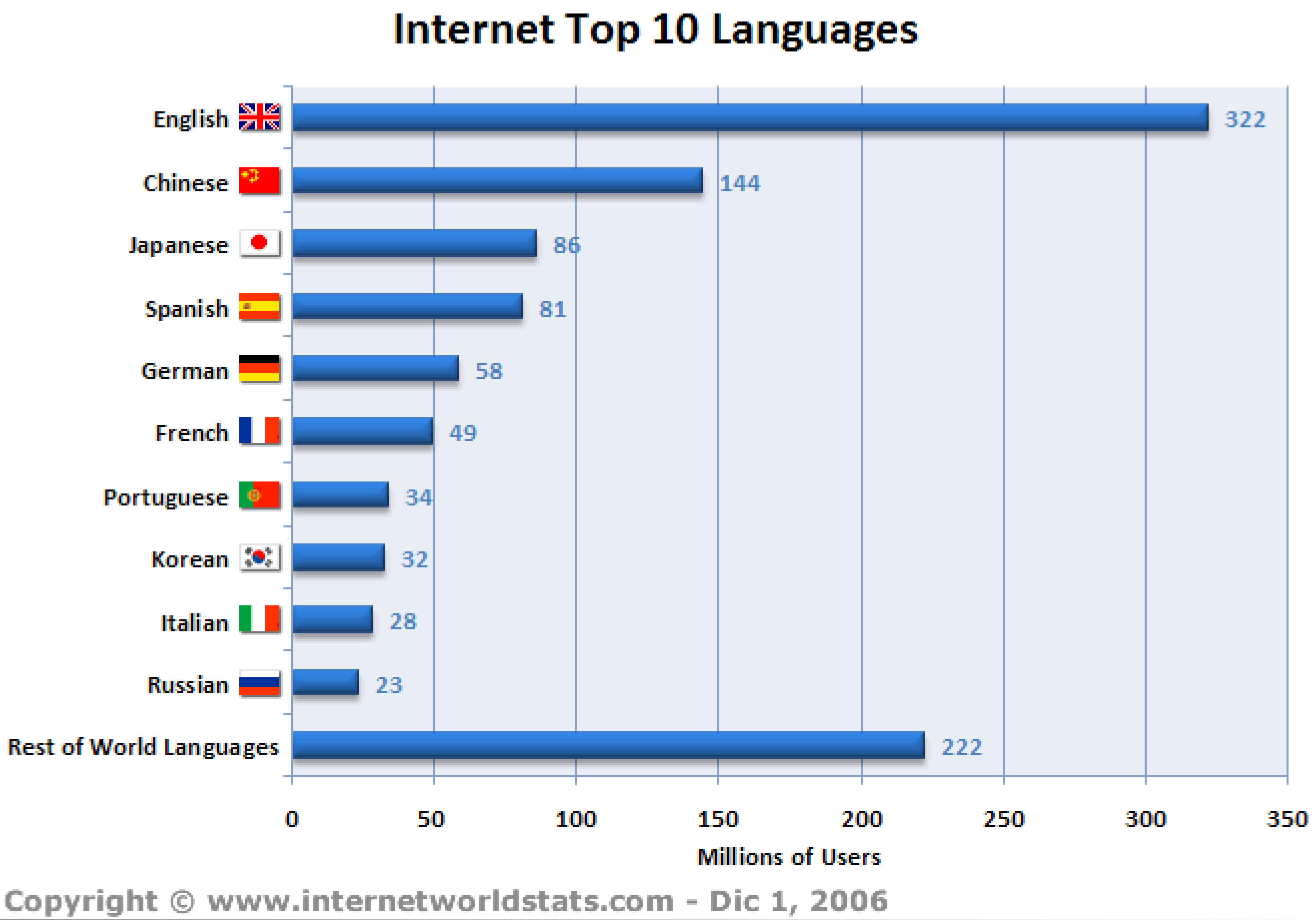
6,809 living languages

(Ethnologue, 14th Edition, Barbara F. Grimes, Editor.
copyright © 2000, SIL International.)

6,809 living languages
(Ethnologue, 14th Edition, Barbara F. Grimes, Editor.
copyright © 2000, SIL International.)

Internet top 10 languages

http://www.internetworldstats.com/stats7.htm



Multilingual systems vs MT systems

- Machine translation systems transform source language into target language

- Multilingual systems
  - Interface technology (display/input)
  - Extract information from multilingual (spoken or written) documents
  - Understand queries in multiple languages

Multilingual systems vs MT systems

- Machine translation systems
  - Systran (about 10mm Euro annual revenue)
  - AltaVista Babelfish (11 languages)

- Multilingual systems
  - Google search (117 languages)
  - Nuance products (OCR/119 languages, SLT/46 languages) (128mm Q4 06)
  - Document editing systems
  - Spam filters
  - Wikipedia (12 languages)

Multilingual systems vs MT systems

- **MT systems are very interesting**
- **Multilingual systems are important**

Multilingual language processing

- Multilingual data mining
  - Non parallel corpora
  - Lexicon extraction
  - NER
  - WSD
  - Dictionary compilation
- Multilingual IR
  - summarization
  - Cross-lingual retrieval
  - Mixed language query processing
- Multilingual linguistic processing
  - POS tagging/chunking
  - Syntactic parsing
  - Semantic parsing
  - Semantic network
- Multilingual speech processing
  - Acoustic modeling
  - Language modeling
  - TTS
  - Pronunciation modeling

Multilingual language processing

- Multilingual data mining
  - Non parallel corpora
  - Lexicon extraction
  - NER
  - WSD
  - Dictionary compilation
- Multilingual IR
  - summarization
  - Cross-lingual retrieval
  - Mixed language query processing
- Multilingual linguistic processing
  - POS tagging/chunking
  - Syntactic parsing
  - Semantic parsing
  - Semantic network
- Multilingual speech processing
  - Acoustic modeling
  - Language modeling
  - TTS
  - Pronunciation modeling

Comparability of bilingual corpora

Non-comparability →

Parallel Corpora
- Translated texts
- No insertion/ deletion

Noisy parallel corpora
- Translated texts
- With insertion/ deletion

Comparable corpora
- Same-topic (in-topic) or time-aligned documents
- May or may not be translated

Very-non-parallel corpora
- Both in-topic and **off-topic** documents
- **Not translated**

Comparability of bilingual corpora

Non-comparability →

Parallel Corpora

Hong Kong Laws
313k sentence pairs in Chinese & English

Noisy parallel corpora

Hong Kong News
Xinhua News;
Technical manuals
with different examples

Comparable corpora

Newspapers in Chinese & English from around the same publication dates

Very-non-parallel corpora

TDT3 data
Chinese & English news articles from different sources and publication dates

Alignment & lexicon mining methods

Non-comparability →

Parallel Corpora	Noisy parallel corpora	Comparable corpora	**Very-non-parallel corpora**
Manning and Schűtze (1999), Wu (2000), and Veronis (2002)	Fung & McKeown, 1995; Kikui 1999; Zhao & Vogel, 2002;	Rapp, 1995; Fung & MckKeown, 1997; Grefenstette, 1998; Fung & Lo, 1998; Kaji, 2003; Munteanu & Marcu, 2004 Sproat et al. 2006	Fung & Cheung,2004; Wu & Fung 2005; Munteanu & Marcu 2006

K-vec (parallel corpora)

Fung & Church 1994

- Parallel texts are divided into K parts
- For each word, its occurrence in the i-th segment is either 1 or 0
- Measure the co-occurrence of two bilingual K-vecs by mutual information and t-score
- Extract the highest ranking pairs and anchor the corpus
- Find other bilingual word pairs in between the anchor points

Table 8: K=100

	pêches	
fisheries	5	1
	0	94

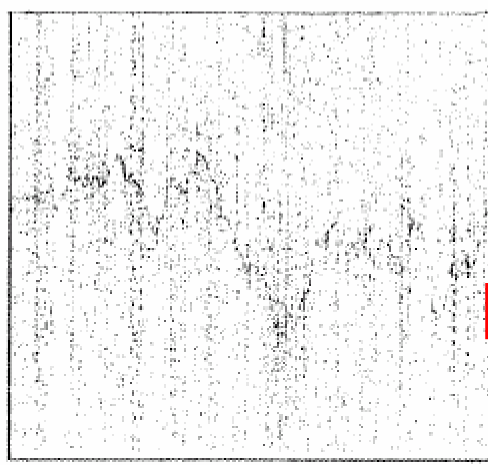


Table 9: K-vec results

	French	English
3.2	Beauce	Beauce
3.2	Comeau	Comeau
3.2	1981	1981
3.0	Richmond	Richmond
3.0	Rail	VIA
3.0	pêches	Fisheries
2.8	Deans	Deans
2.8	Prud	Prud
2.8	Prud	homme
2.7	acheteur	Limited
2.7	Communications	Communications
2.7	MacDonald	MacDonald
2.6	Mazankowski	Mazankowski
2.5	croisière	nuclear

DK-vec (noisy parallel corpora)

Fung & McKeown 1995

- For each word, find its position and recency in the text (DK-vec)
- Measure the co-occurrence of two bilingual DK-vecs by DTW score
- Extract the highest ranking pairs and anchor the corpus
- Find other bilingual word pairs in between the anchor points using K-vecs

DK-vec (Fung&McKeown 1995)

DK-vec

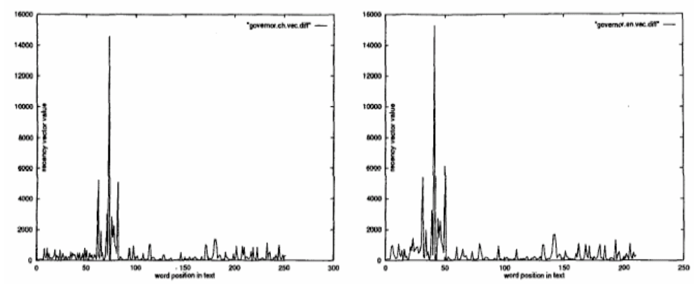
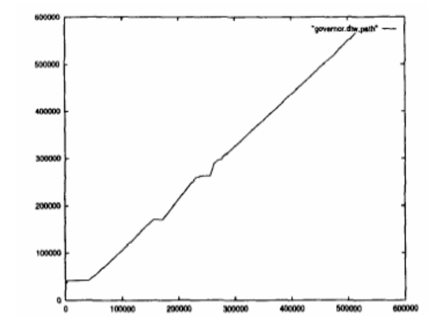


Figure 1: Positional difference signals showing similarity between *Governor* in English and Chinese

DTW path



Alignment path construction

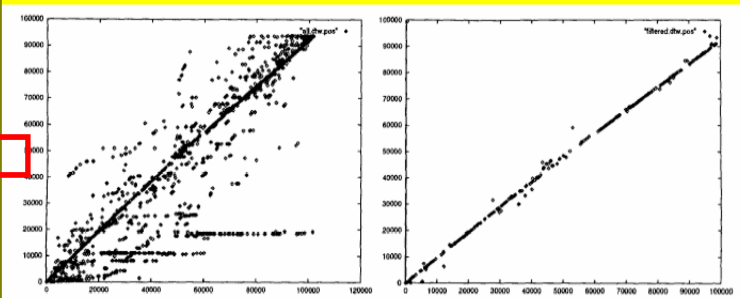


Figure 3: DTW path reconstruction output and the anchor points obtained after filtering

K-vec

$$
\begin{aligned}
m &= \log_2 \frac{\Pr(V1, V2)}{\Pr(V1)\Pr(V2)} \\
\Pr(V1) &= \frac{\mathrm{freq}(V1[i] = 1)}{L} \\
\Pr(V2) &= \frac{\mathrm{freq}(V2[i] = 1)}{L} \\
\Pr(V1, V2) &= \frac{\mathrm{freq}(V1[i] = V2[i] = 1)}{L} \\
\text{where } L &= \dim(V1) = \dim(V2)
\end{aligned}
$$

Terminology translation from noisy parallel corpora

lexicons	total word pairs	correct pairs			accuracy		
		E1	E2	E3	E1	E2	E3
primary(1)	128	101	107	90	78.9%	83.6%	70.3%
secondary(1)	533	352	388	382	66.0%	72.8%	71.7%
total(1)	661	453	495	472	68.5%	74.9%	71.4%
primary(3)	128	112	101	99	87.5%	78.9%	77.3%
secondary(3)	533	401	368	398	75.2%	69.0%	74.7%
total(3)	661	513	469	497	77.6%	71.0%	75.2%

Context-vec (non-parallel corpora)

Fung 1995

Table 2. Words in the context of *flu*/流感 are similar.

English	TF	Chinese		TF
bird	284	事件	(event)	218
virus	49	病毒	(virus)	217
people	45	政府	(establishment)	207
Sydney	38	感染	(contraction)	153
scare	32	表示	(denote)	153
spread	19	沒有	(doesn't_exist)	134
deadly	19	病人	(invalid)	106
government	16	專家	(consultancy)	100
China	14	部門	(branch)	96
new	13	染上	(catch)	93
crisis	13	醫院	(hospital)	92
outbreak	12	情況	(circumstance)	90
hospital	12	處理	(deal_with)	89
chickens	9	醫生	(doctor)	49
spreading	8	染上	(infected)	47
prevent	8	醫院	(hospital)	44
crisis	8	沒有	(no)	42
health	8	政府	(government)	41

$$S(W_c, W_e) = \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\Sigma_{i=1}^{t} w_{ic}^2 \times \Sigma_{i=1}^{t} w_{ie}^2}}$$

where $w_{ic} = TF_{ic} \times IDF_i$

$w_{ie} = TF_{ie} \times IDF_i$

30%-78% accuracy

Other word signature features from non-parallel corpora (Fung & McKeown 1995, Fung 1996,)

seed word	corr1(text1)	seed word	corr1(text2)
amount	1083.35	amount	1083.35
July	695.58	offered	646.30
offered	646.30	preferred	551.50
Canadian	596.42	July	695.58
preferred	551.50	June	393.14
June	393.14	exchange	387.16
exchange	387.16	issue	373.80
issue	373.80	notes	229.45
notes	229.45	gas	158.60
gas	158.60	Capital	157.64

Figure 2: Most correlated seed words with *debentures*

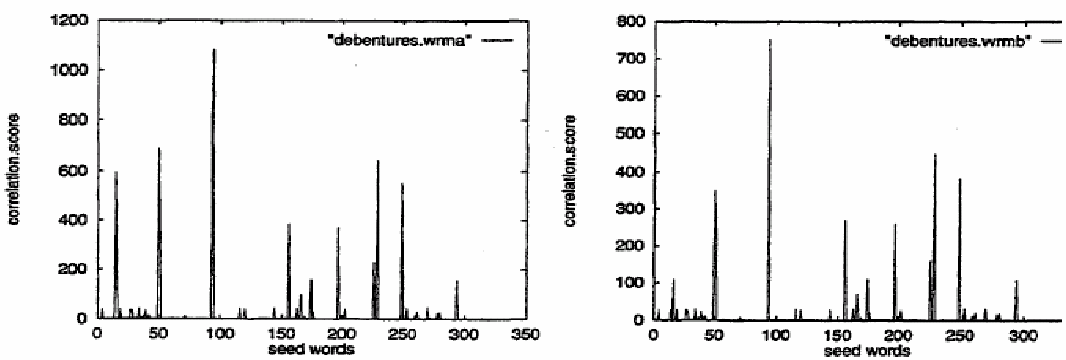


Figure 3: Word relation matrix for *debenture* in both texts

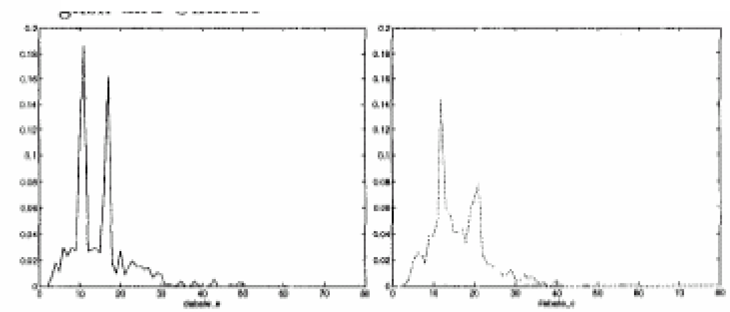


Figure 6: Normalized histogram of *debate* in English and Chinese

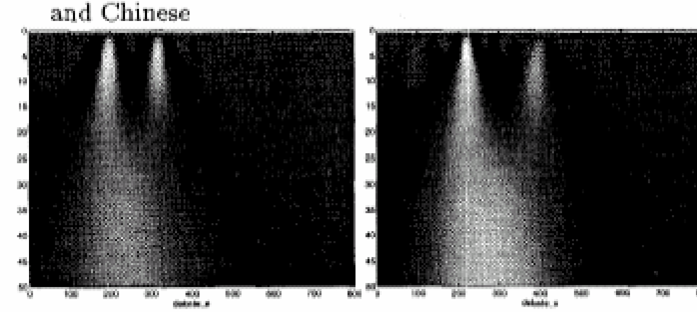
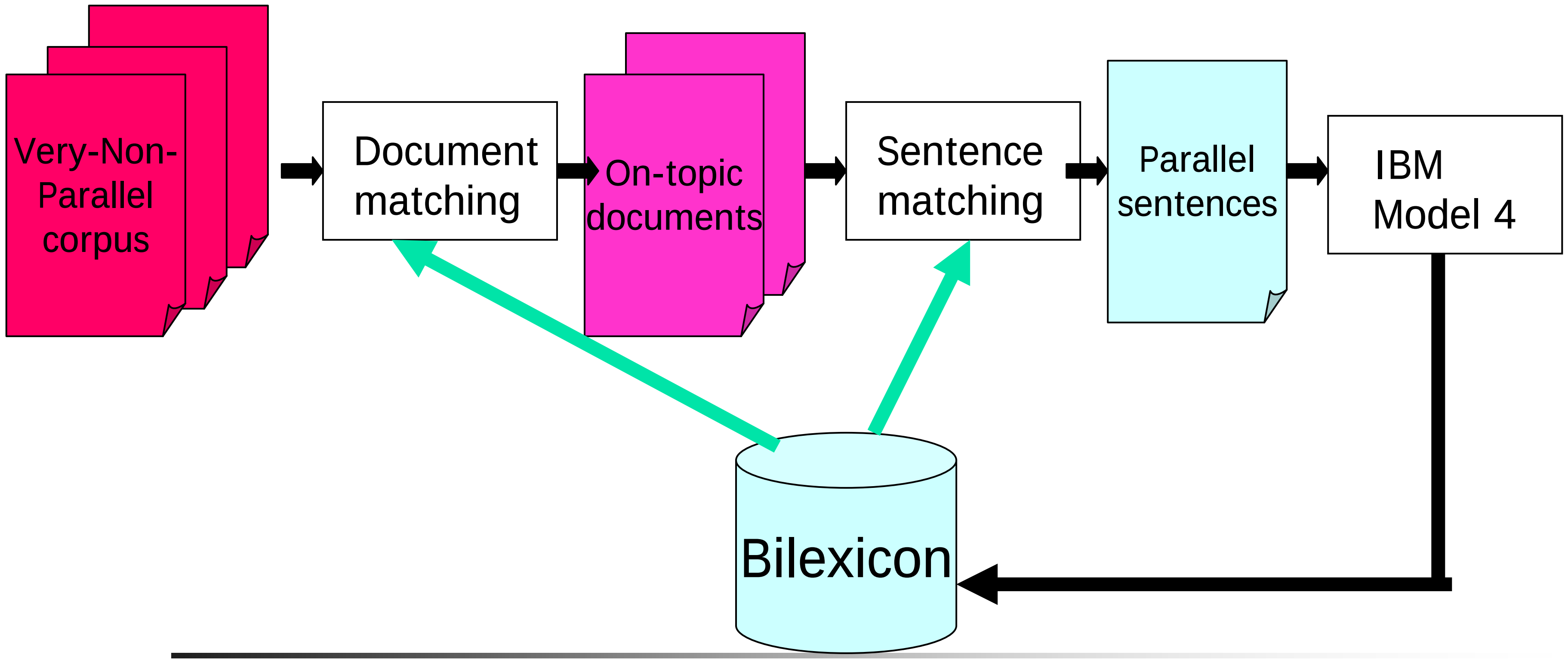
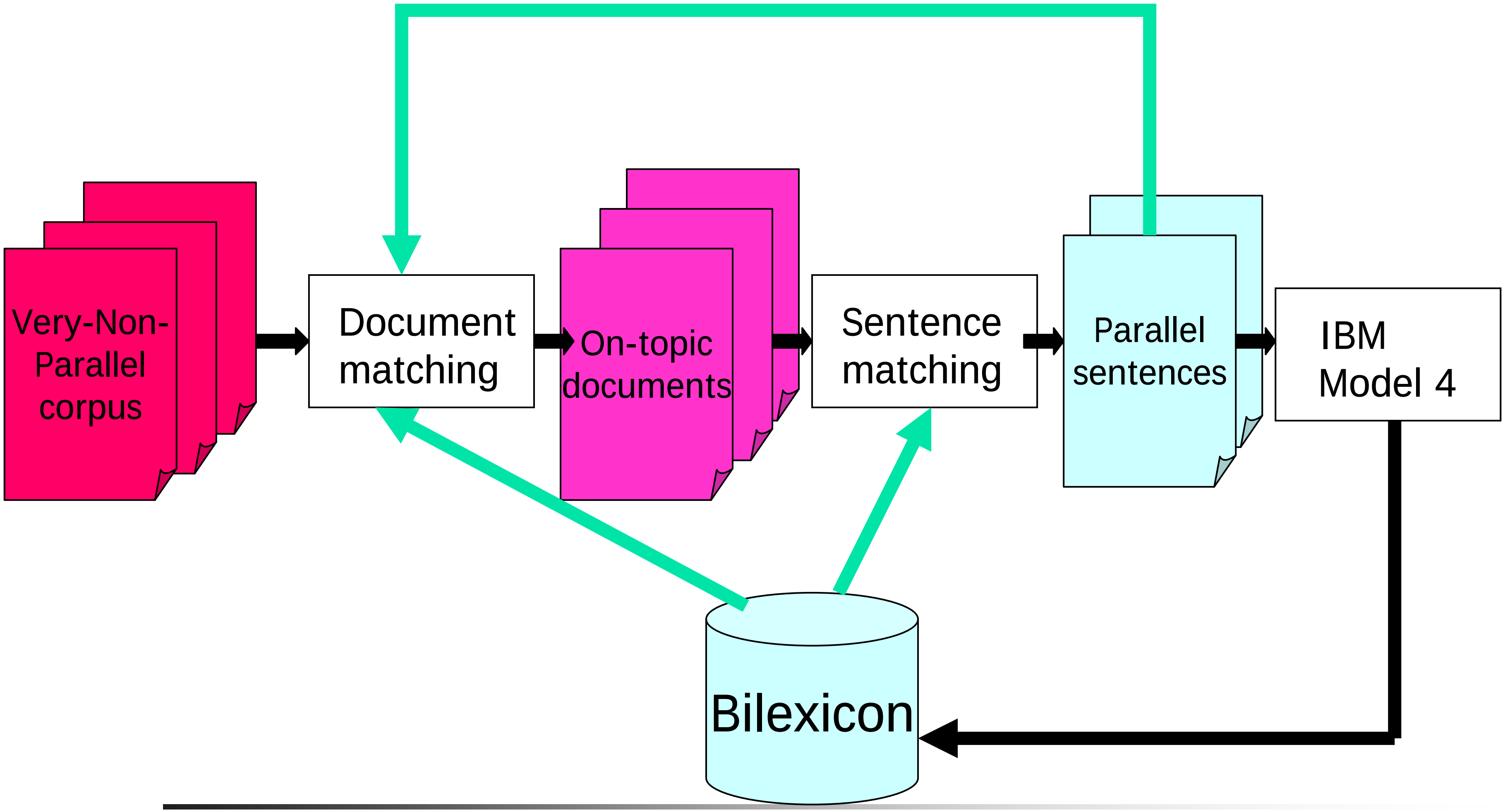


Figure 7: Space-frequency plots of *debate* in English and Chinese

Finding parallel sentences and bilingual lexicon from very non-parallel corpora
(Fung & Cheung 2004)



Finding parallel sentences and bilingual lexicon from very non-parallel corpora
(Fung & Cheung 2004)



Mining strictly parallel sentences with ITG structures (Wu & Fung 2005)

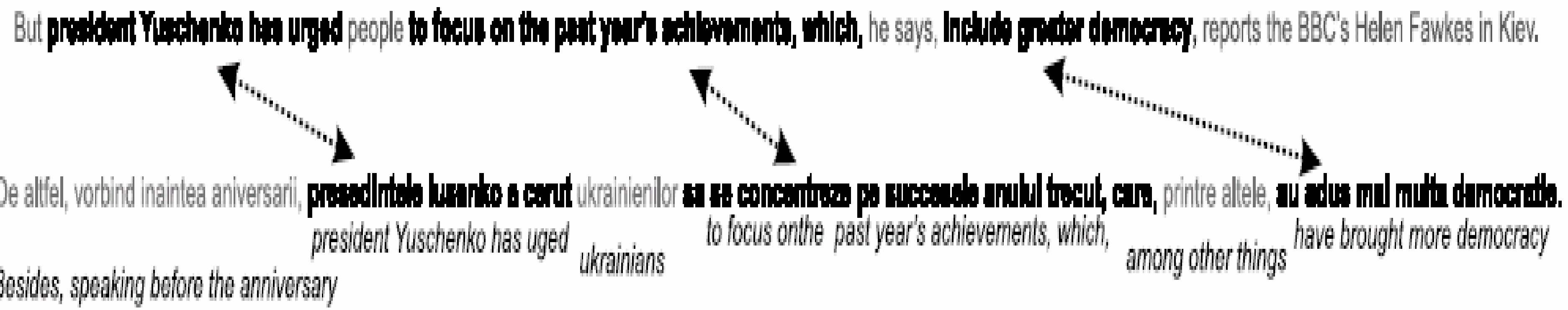
- Candidate generation using bootstrapping and EM
- ITG scoring to find strictly parallel sentences with nested inversions:

It is time to break the silence.
现在 呢 ，是 打破 沉默 的 时候 了 。
(*Now* topical , *is break silence* genitive *time* aspectual .)

I think that's what people were saying tonight.
我 认为 这 是 人们 今晚 所 说 的话 。
(*I think this is people today by say* genitive *words* .)

If the suspects are convicted, they will serve their time in Scotland.
如果 两 名 嫌疑 人 被 判 有罪 ，就 得 在 苏格兰 服刑 。
(*If two* classifier *suspected person* bei-particle *sentence guilty, then must in Scot-land serve time* .)

Extracting parallel sub-sentential fragments from non-parallel corpora
(Munteanu & Marcu 2006)



Extracting parallel sub-sentential fragments from non-parallel corpora
(Munteanu & Marcu 2006)

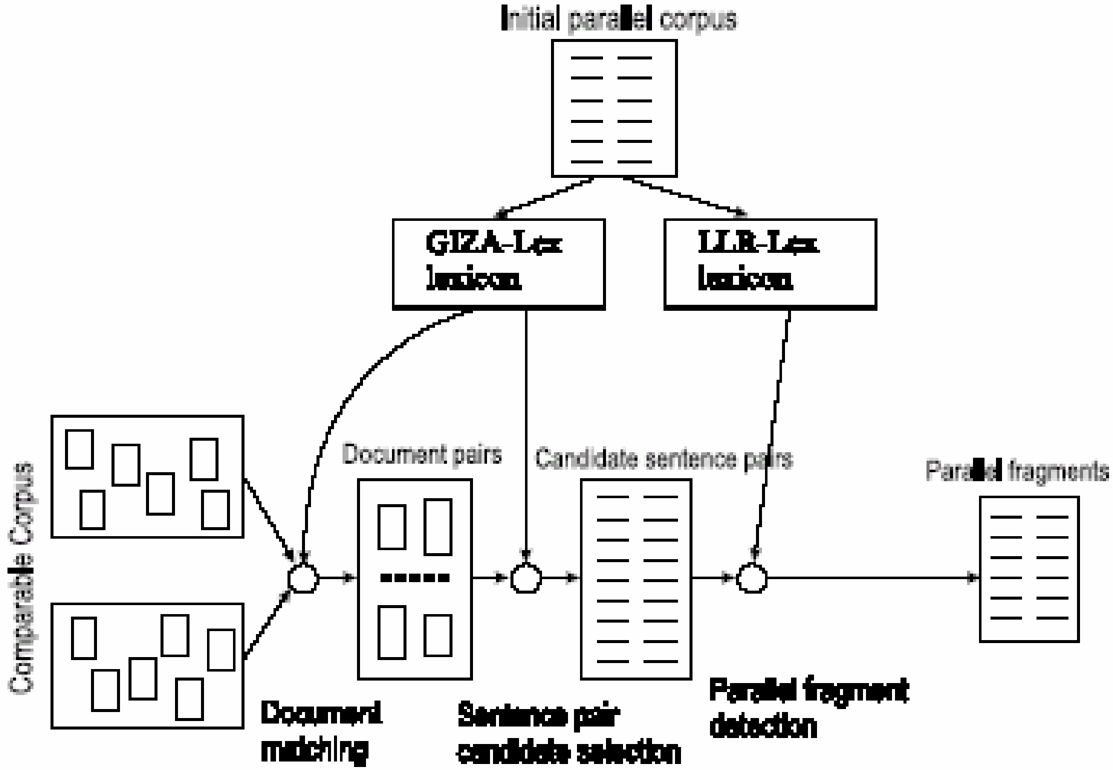
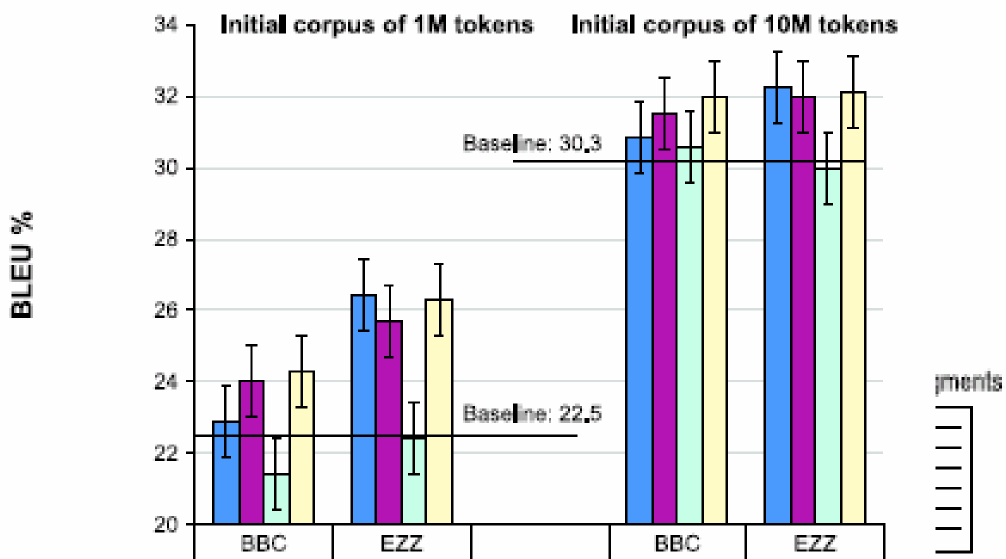


Figure 3: A Parallel Fragment Extraction System

Extracting parallel sub-sentential fragments from non-parallel corpora
(Munteanu & Marcu 2006)



	BBC	EZZ		BBC	EZZ
Plus SentenceExtract	22.92	26.45		30.85	32.21
Plus FragmentExtract	24.04	25.72		31.51	31.98
Plus Fragment-noLLR	21.45	22.45		30.59	29.95
Plus Fragment+Sentence	24.28	26.28		31.99	32.1

Figure 3: A Parallel Fragment Extraction System

Language-independent named entities recognition

- Finding the person name, organization name, location, time, in multiple languages
- Segmentation, part-of-speech tagging, named entities identification and classification

- The CoNNL 2003 shared task: language independent methods for NER
- Maximum Entropy (A Borthwick et al 1999)
- Combining morphological and contextual information (S Cucerzan, D Yarowsky EMNLP 1999)
- HMM based chunk tagger (Zhou & Su 2001)
- Ergodic HMM and statistical bi-gram model (Bikel et al. 1999)
- Multilingual entity mention and tracking (Florian et al. 2004)

Multilingual word sense disambiguation

- Identifying or disambiguating the correct sense of a word
- Sense labeling
- Translation disambiguation

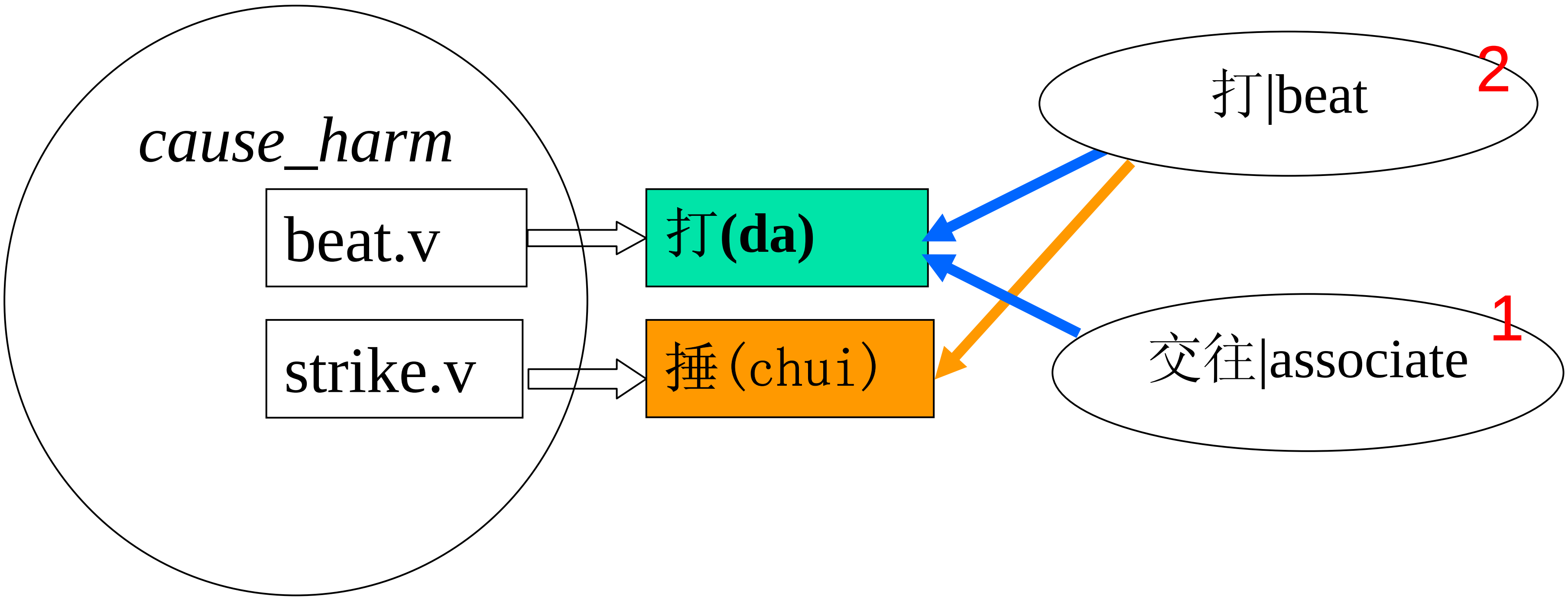
- Using context words and discourse surrounding the source word and use methods ranging from
  - Bilingual bootstrapping (Li & Li 2003),
  - EM iterations (Cao and Li, 2002; Koehn and Knight 2000),
  - and the cohesive relation between the source sentence and translation candidates (Fung et al. 1999; Kikui 1999).

Word sense translation (Fung & Chen 2004)

- Identify and quantify sense mapping between semantic structures across language pairs
- Use existing annotated resources such as

  Framenet,Propbank,Hownet, Wordnet, etc.

Word sense translation:
Induction from bilingual lexicon and FrameNet/HowNet



(1) categories/frames with a large number of translated words should map to each other
(2) lexical entries under aligned categories/frames should map to each other

Example word sense translations

```
tie.n,clothing -> 襻.n,part|部件
tie.v,cause_confinement -> 拘束.v,restrain|制止
tie.v,cognitive_connection -> 联结.v,connect|连接
```

```
make.n,type -> 性质.n,attribute|属性
make.v,building -> 建造.v,build|建造
make.v,causation -> 令.v,CauseToDo|使动
```

```
roll.v,body-movement -> 摇动.v,wave|摆动
roll.v,mass_motion -> 翻滚.v,roll|滚
roll.v,reshaping -> 卷.v,FormChange|形变
```

```
feel.n,sensation -> 手感.n,experience|感受
feel.v,perception_active -> 觉得.v,perception|感知
feel.v,seeking -> 摸.v,LookFor|寻
```

Translation accuracies of 11 most ambiguous words in FrameNet

English word	Number of frames/senses in FrameNet	Sense translation accuracy
tie	8	64%
make	7	100%
roll	6	55%
feel	6	88%
can	5	81%
run	5	100%
shower	5	100%
burn	5	91%
pack	5	85%
drop	5	76%
look	5	64%
Average	**5.6**	**82%**

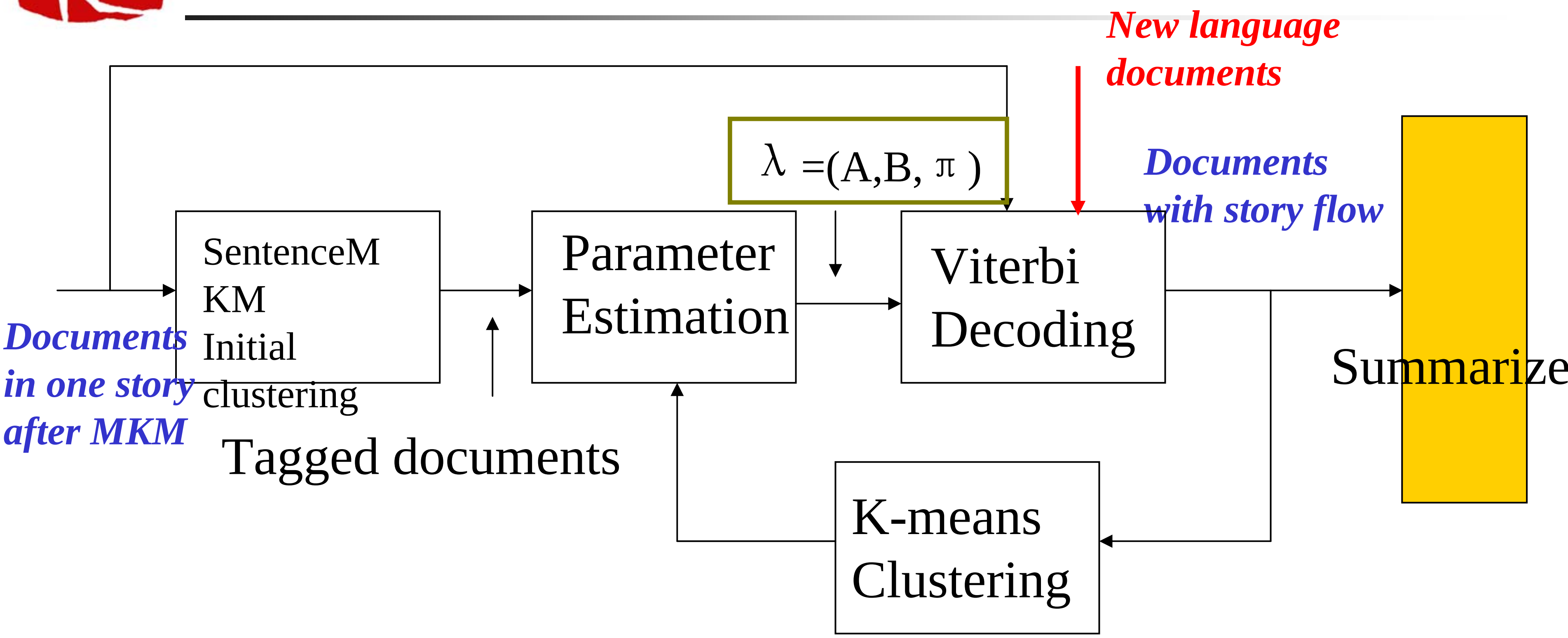
Multilingual language processing

- Multilingual data mining
  - Non parallel corpora
  - Lexicon extraction
  - NER
  - WSD
  - Dictionary compilation
- Multilingual IR
  - summarization
  - Cross-lingual retrieval
  - Mixed language query processing
- Multilingual linguistic processing
  - POS tagging/chunking
  - Syntactic parsing
  - Semantic parsing
  - Semantic network
- Multilingual speech processing
  - Acoustic modeling
  - Language modeling
  - TTS
  - Pronunciation modeling

Multilingual Summarization

- One of the most robust and domain-independent summarization approaches is extraction-based summarization (Mani (1999)).
- Salient sentences are automatically extracted to form a summary directly (Kupiec et. al, (1995), Myaeng & Jang (1999), Jing et. al, (2000), Nomoto & Matsumoto (2001,2002), Zha (2002), Osborne (2002)), or followed by a synthesis stage to generate a more natural summary (McKeown & Radev (1999), Hovy & Lin (1999)).
- Existing multilingual summarization systems (e.g. Radev (2002)) are extensions of a monolingual summarizer with bilingual summaries. In their paradigm, summarization is achieved by extracting salient sentences from monolingual documents, and the multilingual summary is presented as aligned sentences in another language.

One Story One Flow: Hidden Markov Story Models for Multilingual Multi-document Summarization
(Fung et al 2003, Fung & Ngai 2005)



One Story One Flow: Hidden Markov Story Models for Multilingual Multi-document Summarization

The SKM algorithm:

- **Initialization**: All sentences in documents of the same story are clustered using MKM. An initial K states are estimated. Each sentence is given its initial state label. Initial state transitions are counted.
- **(Re-)clustering**: Sentence vectors with their state labels are repartitioned into K clusters (K is obtained from the MKM step previously) using the K-means algorithm:
  - Assign vectors closest to each centroid to its cluster;
  - Update centroid using all vectors assigned to each cluster;
- This step is iterated until the clusters stabilize.
- **(Re-)estimation of probabilities**: The centroids of each cluster are estimated. Update emission probabilities from the new clusters.
- **(Re-)classification by decoding**: the updated set of model parameters from step 2 are used to rescore the (unlabeled) training documents into sequences of story states given sentences, using Viterbi decoding. Update state transitions from this output.
- **Iteration**: Stop if convergence conditions are met, else repeat steps 2-4.

Mixed language query understanding

- Monolingual
  - all words in one language
- Multilingual
  - one language in one query
  - different query in different languages
- Mixed language
  - two (or more?) languages in one utterance/sentence
  - prevalent in Asian languages (e.g. Chinese + English)
  - one utterance consists of words in the primary language and the secondary language

- **Generating translation candidates**
  - **from online dictionary**
- **Weighting translation candidates**
  - **from statistical models**
- **Pruning translation candidates**
  - **Uses contextual words as disambiguating features.**

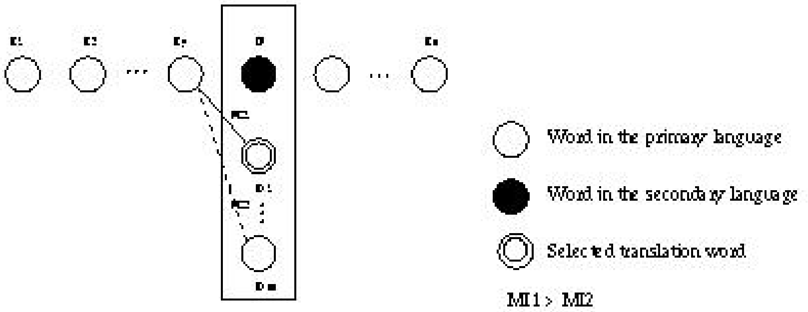


Figure 1: The neighboring word as disambiguating feature

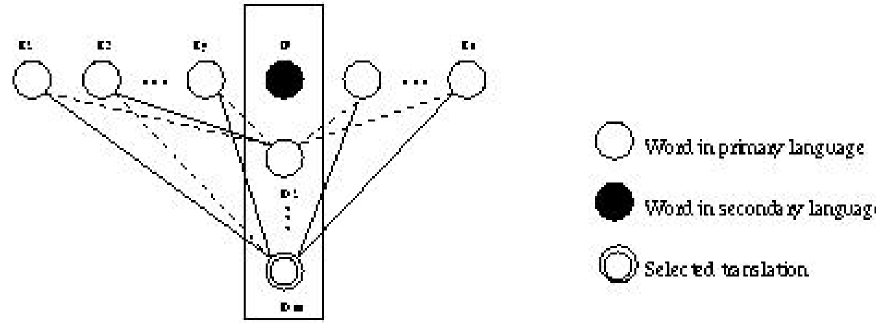


Figure 2: Voting for the best translation

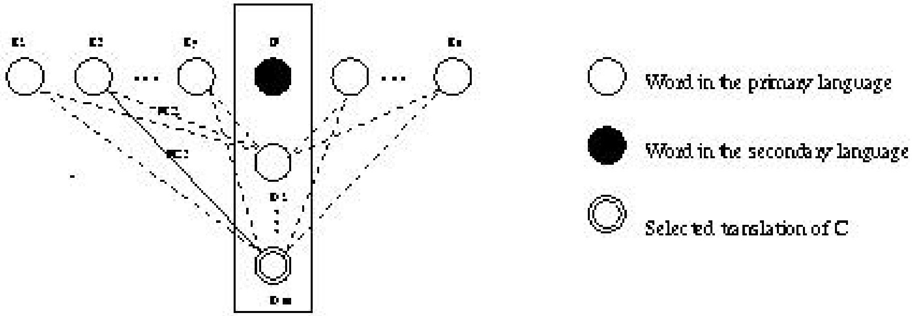


Figure 3: The best contextual word as disambiguating feature

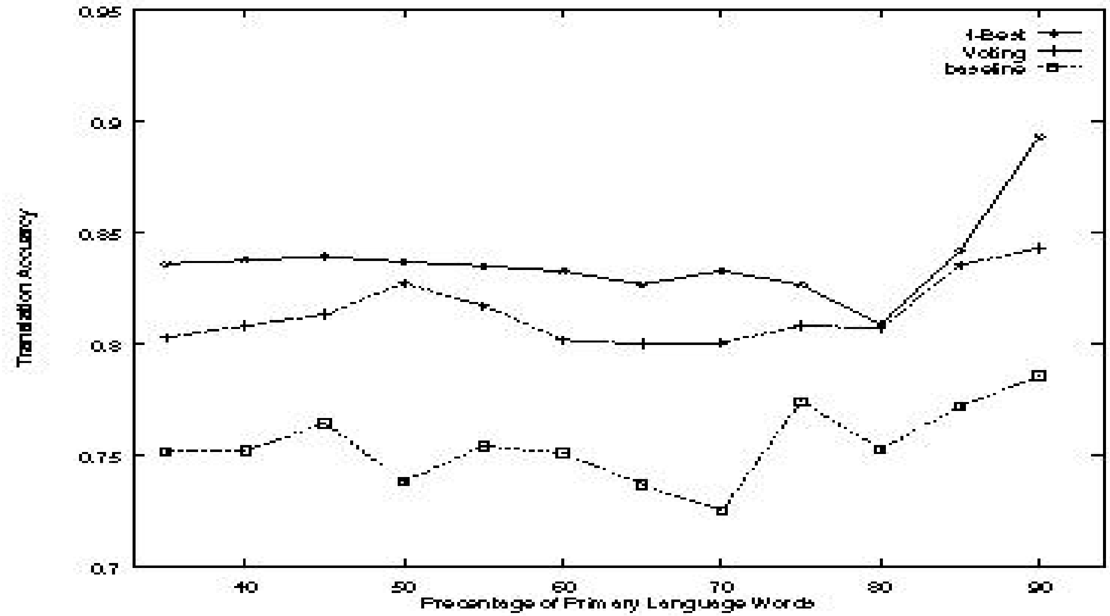


Figure 4: 1-best is the most discriminating feature

- Unsupervised training without bilingual or mixed language training data
- Monolingual queries can be generated in both primary and secondary languages for cross-language IR

The Future of Multilingual IR

- Summarization beyond sentence extraction (Evans 2006)
  - Information analysis
  - Factual information analysis
  - Automatic controversy identification
  - Opinion identification
- More language independent tools for multiple languages
- Multilingual lexicons and ontologies
- Focus on *differences* rather than shared information between languages and cultures

Multilingual language processing

- Multilingual data mining
  - Non parallel corpora
  - Lexicon extraction
  - NER
  - WSD
  - Dictionary compilation
- Multilingual IR
  - summarization
  - Cross-lingual retrieval
  - Mixed language query processing
- Multilingual linguistic processing
  - POS tagging/chunking
  - Syntactic parsing
  - Semantic parsing
  - Semantic network
- Multilingual speech processing
  - Acoustic modeling
  - Language modeling
  - TTS
  - Pronunciation modeling

Multilingual linguistic processing

- Multilingual POS tagging
  - By coercion (Fung & Wu, 1995) and projection (Yarowsky & Ngai 2001)
- Multilingual syntactic parsing
  - Lexicalized statistical parser with language packages (D. Bikel EMNLP 2004)
  - Language-specific packages include Treebank processing, preprocessing, word features
- Multilingual semantic parsing
  - SVM-based statistical parser in English (Pradhan et al. 2004-5) and in Chinese (Fung et al. this workshop)
  - ME-based Chinese and English parsers (Fung et al 2004, Xu et Palmer 2005)
- Semantic network mapping
  - Wordnet/HowNet (Carpuat, Fung & Ngai, 2005)
  - FrameNet/HowNet (Fung & Chen 2004/5), Propbanks (Fung et al. this workshop)

Multilingual language processing

- Multilingual data mining
  - Non parallel corpora
  - Lexicon extraction
  - NER
  - WSD
  - Dictionary compilation
- Multilingual IR
  - summarization
  - Cross-lingual retrieval
  - Mixed language query processing
- Multilingual linguistic processing
  - POS tagging/chunking
  - Syntactic parsing
  - Semantic parsing
  - Semantic network
- Multilingual speech processing
  - Acoustic modeling
  - Language modeling
  - TTS
  - Pronunciation modeling

Multilingual speech processing

- Multilingual acoustic modeling
  - Phonetic mapping (Fung et al 1997)
  - Phone model adaptation (Ma & Fung 1998)
  - (Fung & Liu 2000, Shultz & Katrin Kirchhoff 2006)
- Multi-accent pronunciation modeling
  - (Fung & Liu 2000, Liu & Fung 2006)
- Multilingual language modeling (Fung & Lo 1999; Sproat 2004)
- Multilingual dialog systems (Fung et al.;Meng et al.)

Reference:

- Tanja Schultz & Katrin Kirchhoff, *Multilingual Speech Processing*, Academic Press, 2006.

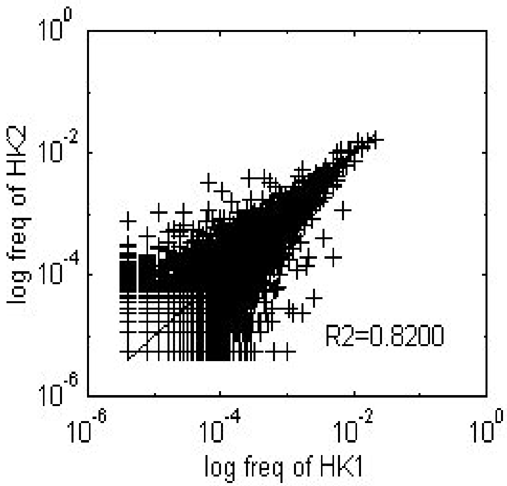
Mandarin/English phonetic mapping

Man	Eng	Man	Eng	Man	Eng	Man	Eng	Man	Eng
a	aa	eng	aa ng	ing	iy ng	q	ch	uen	er n
ai	ae	er	aa	iong	uw ng	r	y	ueng	aa ng
an	ae ng	f	f	iou	ow	s	s	uo	ao
ang	aa ng	g	g	j	y	sh	sh	ü	iy
ao	aw	g	hh	k	k	t	hh	üan	ae ng
b	b	i	iy	l	y	u	uw	üe	ey
c	th	ia	aa	m	m	ua	aa	ün	ey ng
ch	ch	ian	ae ng	n	y	uai	ay	w	w
d	b	iang	aa ng	o	ao	uan	ay ng	x	s
e	aa	iao	aw	ong	ow ng	uang	aa ng	y	y
ei	ey	ie	ey	ou	ow	uei	ey	z	th
en	ae n	in	iy ng	p	p	ue	er n	zh	jh

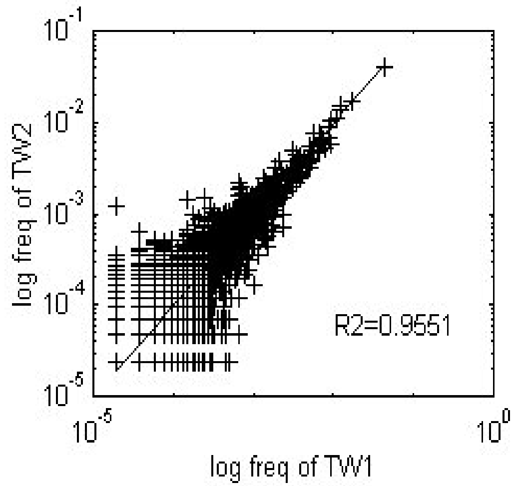
Cantonese/English phonetic mapping

Can	Eng	Can	Eng	Can	Eng	Can	Eng	Can	Eng
aa	aa	at	ah t	f	f	l	l	p	p
aai	ay	au	aw	g	g	m	m	s	s
aak	aa k	b	b	gw	g w	n	n	t	t
aam	aa m	c	ch	h	hh	ng	ng	u	uw
aan	aa n	d	d	i	iy	o	ao	ue	uw
aang	aa ng	e	ea	ik	ih k	oe	er	uen	uw n
aap	aa p	ei	ey	im	iy m	oei	uh	uet	uw t
aat	aa t	ek	ea k	in	iy n	oek	er ng	ui	uh
aau	aw	em	ea m	ing	ih ng	oeng	oy	uk	uh k
ai	ay	eng	ea ng	ip	ih p	oi	ay	un	uw n
ak	ah k	eoi	uw	it	ih t	ok	ao k	ue	uw
am	ah m	eon	uh n	iu	iy	on	ao n	ung	uh ng
an	an	eot	uh t	j	y	ong	ao ng	ut	uw t
ang	ah ng	ep	ea p	k	k	ot	ao t	z	jh
ap	ah p	eu	uw	kw	kw	ou	ow		

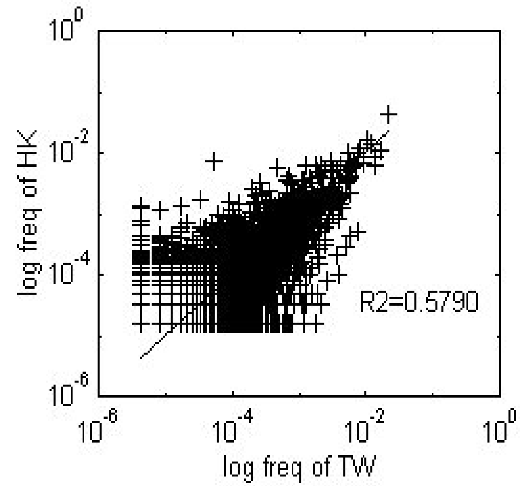
Cantonese Mandarin LM



Cantonese/Cantonese



Mandarin/Mandarin



Cantonese/Mandarin

Cantonese Mandarin LM

- Extraction of Cantonese content words and terms
- Extraction of Cantonese filler/colloquial words and terms
- Augment the segmentation lexicon with these terms
- Interpolation of language model

$$P_{M,C}(w_i \mid h) = \lambda_i P_M(w_i \mid h) + (1 - \lambda_i) P_C(w_i \mid h),$$

Table 10. The interpolated language model gives the best perplexity measure on a Cantonese colloquial test data set. The lexicon for all language models is the augmented lexicon with 43,234 entries.

Language model	Language	Training data	Perplexity
LM1	Written Chinese, Mandarin.	Ming Pao newspaper text, 260Mb, 5.1M words	2586.28
LM2	Colloquial Chinese, Cantonese.	Hong Kong Newsgroup Corpus, 4Mb, 121K words	364.27
LM3	Concatenated database	Newspaper and News Group Corpus	1755.66
LM4 $\lambda_i = 0.48$	**Interpolated model**	**Interpolated model**	**307.882**

Multilingual Applications

- Spoken language understanding & generation
- Information retrieval (legal/medical/education)
- Table understanding
- Internet businesses
- Search
- Spam filters
- User Generated Content

TOP 20 COUNTRIES WITH HIGHEST NUMBER OF INTERNET USERS

#	Country or Region	Internet Users, Latest Data	Population (2006 Est.)	Internet Penetration	Source and Date of Latest Data	% Users of World
1	United States	209,024,921	299,093,237	69.9 %	Nielsen//NR Oct/06	19.4 %
2	China	123,000,000	1,306,724,067	9.4 %	CNNIC June/06	11.4 %
3	Japan	86,300,000	128,389,000	67.2 %	eTForecasts Dec/05	8.0 %
4	Germany	50,616,207	82,515,988	61.3 %	Nielsen//NR Aug/06	4.7 %
5	India	40,000,000	1,112,225,812	3.6 %	IWS Nov/06	3.7 %
6	United Kingdom	37,600,000	60,139,274	62.5 %	ITU Sept/06	3.5 %
7	Korea (South)	33,900,000	50,633,265	67.0 %	eTForecast Dec/05	3.1 %
8	Italy	30,763,848	59,215,261	52.0 %	Nielsen//NR Oct/06	2.9 %
9	France	29,521,451	61,004,840	48.4 %	Nielsen//NR Aug/06	2.7 %
10	Brazil	25,900,000	184,284,898	14.1 %	eTForcasts Dec/05	2.4 %
11	Russia	23,700,000	143,682,757	16.5 %	eTForcasts Dec/05	2.2 %
12	Canada	21,900,000	32,251,238	67.9 %	eTForcasts Dec/05	2.0 %
13	Mexico	20,200,000	105,149,952	19.2 %	AMIPCI Oct/06	1.9 %
14	Spain	19,204,771	44,351,186	43.3 %	Nielsen//NR Oct/06	1.8 %
15	Indonesia	18,000,000	221,900,701	8.1 %	eTForcasts Dec/05	1.7 %
16	Turkey	16,000,000	74,709,412	21.4 %	ITU Sept./06	1.5 %
17	Australia	14,663,522	20,750,052	70.7 %	Nielsen//NR Aug/06	1.4 %
18	Taiwan	13,800,000	22,896,488	60.3 %	C.I.Almanac Mar/05	1.3 %
19	Poland	11,400,000	38,115,814	29.9 %	Survey Oct./06	1.1 %
20	Netherlands	10,806,328	16,386,216	65.9 %	Nielsen//NR June/04	1.0 %
TOP 20 Countries		836,301,148	4,064,319,458	20.6 %	IWS - Nov. 27/06	77.7 %
Rest of the World		239,902,839	2,435,377,602	9.9 %	IWS - Nov. 27/06	22.3 %
Total World - Users		1,076,203,987	6,499,697,060	16.6 %	IWS - Nov. 27/06	100.0 %

Partial list of references

1. http://www.ee.ust.hk/~pascale/Publications/publications-by-topic.html
2. http://ucdata.berkeley.edu:7101/sigir2006-mlia.htm
3. http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/archive/ACL2003/ws1-mlsum-program.htm
4. http://www.bbn.com/Advanced_Technologies/Data_Indexing_and_Mining/Cross_Lingual_Info.html
5. http://www1.cs.columbia.edu/nlp/
6. http://tangra.si.umich.edu/~radev/publications/
7. T. Schultz and K. Kirchhoff (eds.) *Multilingual Speech Processing* , Elsevier, 2006