Centre for Next Generation Localisation

MWE-sensitive Word Alignment in Factored Translation Model

Tsuyoshi Okita, Andy Way
Dublin City University, Research Workshop of the Israel Science Foundation

National Development Plan 2007 - 2013
science foundation ireland
Dublin City University University College Dublin University of Limerick Trinity College Dublin

Table Of Contents

NTCIR-8 Corpora

- JP-EN Patent Corpus
  - Essentially translated from JP into EN (Quality of EN?).
  - 3.2 million sentence pairs.

	train set	dev set	test set
JP-EN	3,186,284	1,200/2,000	1,251
EN-JP	3,186,284	1,200/2,000	1,119

Table: Parallel corpus size of NTCIR-8

NTCIR-8 Corpora

- Unstructured complex sentences

Japanese: この第２のスライドブロック５のスライド移動によって、弾性糸ＳＹが、第１のスライドブロック４の下流側面と前記第２のスライドブロック５の上流側面との間で、確実に把持されるとともに、前記弾性糸ＳＹは、第２のスライドブロック５の下流側面と下側の固定ブロック６の上流側面とのスライドによって前記カッター刃１０の作用により切断される。

English: due to this slidable movement of the second slide block 5 , the elastic yarn sy is reliably held between the downstream side of the first slide block 4 and the upstream side of the second slide block 5 , and the elastic yarn sy is cut by the operation of the cutter blade 10 due to sliding between the downstream side of the second slide block 5 and the upstream side of the fixed block 6 at the lower side .

NTCIR-8 Corpora

- Translational Omission

  Japanese: 従来、スパッタリング用ターゲット（以下、単にターゲットと略称する）としてはプレーナ型（円板状もしくは角板状）のターゲットが広く使用されている。

  English: conventional sputtering targets extensively in use are of a planer type having a circular or square plate-like shape .

NTCIR-8 Corpora

- Equations in a sentence

  Japanese: 処理７３では、上記の取り込んだ信号Ｖ，θｆ，ｄ／ｄｔ(θｆ）から、目標ヨー角加速度ｄ／ｄｔ(ωＴ）を決定する。

  English: at a process 73 , target yawing angular acceleration d/dt (.omega. .sub.t) is determined based on the fetched signals v , .theta.f and d/dt (.theta.f) .

Centre for Next Generation Localisation

NTCIR-8 Corpora

Reference number

Japanese: この軸受けユニット（44 ）は、本体支柱（４）に固着した上・下部ブラケット（45 ）（46 ）と、この上・下部ブラケット（45 ）（46 ）に挿通した軸セット（47 ）と、軸セット（47 ）と製品容器（２）間を連結するアーム（48 ）とで構成する。
Japanese: 図１に示すガイド５とガイドローラ３はこのような案内を行うものであり、以下にその実施例を説明する。

English: the bearing unit 44 comprises upper and lower brackets 45 and 46 fixed on the body pillar 4 , a shaft set 47 inserted in said upper and lower brackets 45 and 46 , and an arm 48 interconnecting the shaft set 47 and the product container 2.
English: the guide 5 and the guide rollers 3 shown in fig. 1 are designed to provide such guidance , and embodiments thereof will be described hereinunder .

NTCIR-8 Corpora

Many parentheses

Japanese: 次に、直径５ｍｍの多数の空孔（９６ポイント）を有する開閉式の扉１４を開けて、大気中に浮遊している有機ガス１５を基板１０に２４時間吸着させる（ステップＳＡ２）。

English: then , in step 2 of fig. 4 , the cover 14 is opened for 24 hours so that gaseous organic substances 15 floating in an atmosphere are adsorbed to the silicon substrate 10 .

Intrinsic Evaluation (JP-EN)

Systems	BLEU	#OOV
System combination	27.61*	321
HPB-SMT 1	26.86*	314
PB-SMT 1	26.51*	194
Noise reduction (PB-SMT)	24.01	443
PB-SMT 2+	23.91*	316
Preprocessing (PB-SMT)+	23.82	194
HPB-SMT 2	23.30	303
Supertag (ENJU) 1	20.68	430
Supertag (ENJU) 2	18.27	426
System combination (unofficial run)	28.43	331

Table: Intrinsic evaluation results (JP-EN). Noted that we trained over 3,200k training corpus for the systems marked with + and over 600k training corpus for other systems.

Intrinsic Evaluation (EN-JP)

Systems	BLEU
System combination	<u>33.03</u>
HPB-SMT 1	32.50
PB-SMT 1	30.53
PB-SMT 2^+	30.08
Noise reduction	29.53
Preprocessing $(\text{PB-SMT})^+$	27.93
HPB-SMT 2	27.23
Context supertag (Base)	26.83
Context supertag (Superpair)	26.45
Context supertag (CCG)	26.38
Context supertag (LTAG)	26.38
Context supertag (CCG-LTAG)	26.22
Context supertag (POS)	26.21

Table: Intrinsic evaluation results (EN-JP).

Table Of Contents

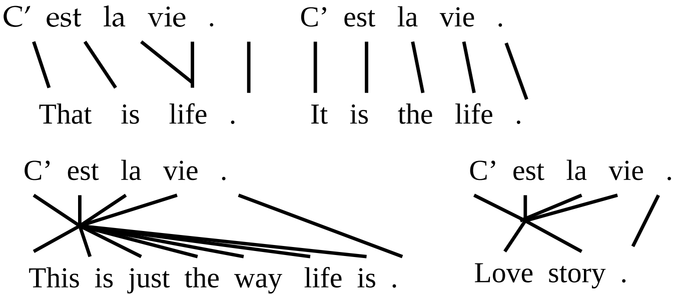
MWE-sensitive Word Alignment onn PB-SMT

approach	IBM disc	reference	word alignment E-step	word alignment M-step	other methods	prior	h-a corp	feat
MWE-sensitive	12H34	[Okita&, 10]	normal	MAP	N.A.	objective	no	MWE
GIZA++	12H3456	[Och&, 02]	normal	ML	N.A.	none	no	no
Berkeley	12H	[Liang&,06]	jointly trained HMM		agreement	none	y-n	no
	12H	[DeNero&,07]	syntax-aware HMM			none	y-n	syn
PostCAT	12H	[Graca&, 08]	constraint	grad asc	agreement	stochastic	no	no
discrimative aln	1d	[Moore, 05]	N.A.	N.A.	bipartite match	N.A.	yes	gen
CRF	1d	[Blunsom&06]	normal	CRF	none	noninform	yes	gen
semi-super	1d	[Fraser, 07]	N.A.	N.A.	EMD	N.A.	yes	gen
joint phrase	B	[Marcu&02]	N.A.		phrase align	none	no	no
delete links	1H34	[Fossum&08]	GIZA++ compatible			none	no	syn
add links	1H34d	[Ma&, 08]	GIZA++ compatible		SVM	none	no	syn
BMWE	1H34	[Lambert&,05]	GIZA++ compatible		N.A.	none	no	no
word lattice	[1H34]	[Dyer&, 08]	GIZA++ compatible		lattice	none	no	no

Table: Word / phrase alignment.

N-to-m Mapping Object Problem (1)

- (Problem in PB-SMT context) N-to-m mapping objects, such as paraphrases, non-literal translations, and multi-word expressions, may appear as both noise and as valid training data for word alignment.



N-to-*m* Mapping Object Problem (2)

Source Language

to my regret i cannot go today .
i am sorry that i cannot visit today .
it is a pity that i cannot go today .
sorry , today i will not be available

Target Language

i am sorry that i cannot visit today .
it is a pity that i cannot go today .
sorry , today i will not be available
to my regret i cannot go today .

GIZA++ alignment results for IBM Model 4

i NULL 0.667
cannot available 0.272
it am 1
is am 1
sorry go 0.667
, go 1
that regret 0.25
cannot regret 0.18
visit regret 1
regret not 1
be pity 1

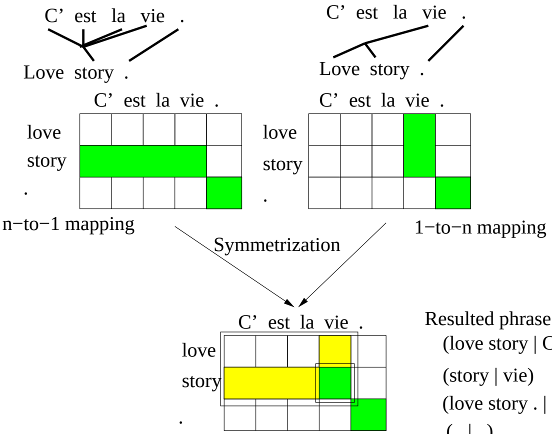
available pity 1
cannot sorry 0.55
go sorry 0.667
am to 1
sorry to 0.33
to , 1
my , 1
will is 1
not is 1
a that 1
pity that 1

today . 1
. . 1
i cannot 0.33
that cannot 0.75

Figure: Paraphrase example: a training corpus consists of four sentence pairs. Results show that only the matching between the colon is correct.

N-to-m Mapping Objects Problem (3)

- How phrase extraction works?



Resulted phrase pairs:

(love story | C' est la vie)

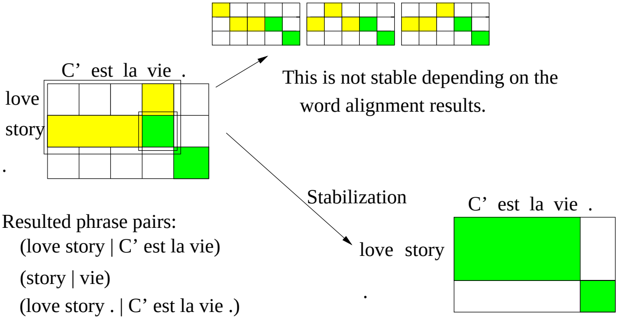
(story | vie)

(love story . | C' est la vie .)

(. | .)

N-to-*m* Mapping Object Problem (4)

- Need to stabilize the temper of word alignment by 'grouping'.



Resulted phrase pairs:

(love story | C' est la vie)

(story | vie)

(love story . | C' est la vie .)

(. | .)

Resulted phrase pairs:

(love story | C' est la vie)

(love story . | C' est la vie .)

(. | .)

MWE-sensitive Word Alignment (1)

Definition (Anchor Word Alignment Problem)

Let $(\breve{e}, \breve{f}) = \{(\breve{e}_1, \breve{f}_1), \ldots, (\breve{e}_n, \breve{f}_n)\}$ be a parallel corpus. By prior knowledge we additionally have knowledge of anchor words $(\hat{e}, \hat{f}) = \{(sent_i, t_{e_1}, t_{f_1}, pos_{e_1}, pos_{f_1}, length_e, length_f), \ldots, (sent_k, t_{e_n}, t_{f_n}, pos_{e_n}, pos_{f_n}, length_e, length_f)\}$ where $sent_i$ denotes sentence ID, pos_{e_i} denotes the position of t_{e_i} in a sentence $\breve{e}_i$, and $length_e$ (and $length_f$) denotes the sentence length of the original sentence which includes e_i. Under a given $(\breve{e}, \breve{f})$ and $(\hat{e}, \hat{f})$, our objective is to obtain word alignments. It is noted that an anchor word may include a phrase pair which forms n-to-m mapping objects.

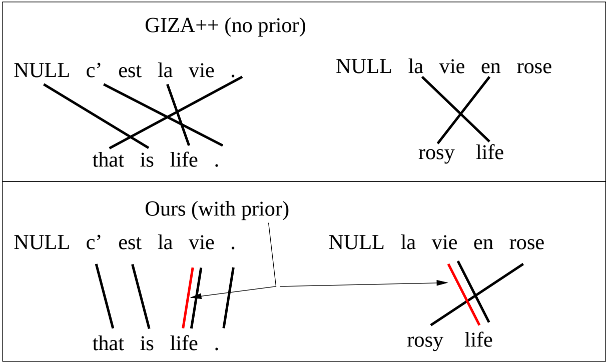
MWE-sensitive Word Alignment (2)

<table>
<tr><th>Statistical MWE extraction method</th></tr>
<tr><td>97|||groupe_socialiste|||socialist_group|||26|||26
101|||monsieur_poettering|||mr_poettering|||1|||4
103|||monsieur_poettering|||mr_poettering|||1|||11
110|||monsieur_poettering|||mr_poettering|||1|||9
117|||explication_de_vote|||explanation_of_vote|||28|||26</td></tr>
<tr><th>Heuristic-based MWE extraction method</th></tr>
<tr><td>28|||the_wheel_2|||車輪_２ ||| 25||| 5
28|||the_primary-side_fixed_armature_13|||１_次_側_固定_電機_子_１_３ ||| 13||| 9
28|||the_secondary-side_rotary_magnet_7|||２_次_側_回転_マグネット_７ ||| 15||| 11</td></tr>
</table>

Table: Example of MWE pairs in Europarl corpus (FR-EN) and NTCIR patent corpus (JP-EN). There are 5 columns for each term: sentence number, source term, target term, source position, and target position.

MWE-sensitive Word Alignment (3)

- We are given two sentence pairs {(that is life . , *c' est la vie .*), (rosy life, *la vie en rose*)} and anchor words $\{(1, \text{life}, \textit{vie}, 3, 4), (2, \text{life}, \textit{vie}, 2, 2)\}$.



MWE-sensitive Word Alignment (4)

pair	GIZA++(no prior)			Ours(with prior)		
EN-FR	fin	ini	prior	fin	ini	prior
is *NULL*	1	.25	0	0	.25	.25
rosy *en*	1	.5	0	0	.5	.2
that .	1	.25	0	0	.25	.25
life *la*	1	.25	0	0	.25	0
. *c'*	1	.25	0	0	.25	.25
that *c'*	0	.25	0	1	.25	.25
is *est*	0	.25	0	1	.25	.25
life *vie*	0	.5	0	1	.5	1
rosy *rose*	0	.25	0	1	.25	.2

Table: The benefit of prior knowledge of anchor words.

2-step Learning Procedure

- Statistical models often aim to learn from frequent examples only, and often overlook less frequent but linguistically important phenomena.
- 2-step learning procedure can capture both frequent and linguistically important phenomena via (Bayesian) priors. Let linguistically important phenomenon be learnable in terms of V and not learnable in terms of W, while let other phenomenon be learnable in terms of W.
  - 1: Learn linguistically important phenomenon in terms of V.
  - 2: Learn the main phenomenon in terms of W with a prior learnt in Step 1.

2-step Learning Procedure

- Prior Knowledge 1: Word alignment with prior knowledge about bilingual terminology (IBM Model 4, hidden variable is alignment function)
  - 1: bilingual terminology: Learn linguistically important phenomenon in terms of word association.
  - 2: Learn word alignment $P(e|f)$ in terms of word pair frequencies (with alignment function as hidden variable) with the prior about an alignment function supplied by bilingual terminology.

$$align(e_i|f_i, T) = \begin{cases} 1 & (e_i = t_i, f_j = t_j) \\ 0 & (e_i = t_i, f_j \neq t_j) \\ 0 & (e_i \neq t_i, f_j = t_j) \\ \text{uniform} & (e_i \neq t_i, f_j \neq t_j) \end{cases}$$

2-step Learning Procedure

Word alignment implemented by EM algorithm: to replace the M-step with Maximum A Posteriori estimate ($p(t) = align(e|f)$ denotes a prior):

$$\begin{aligned} \mathbf{E}^{\mathbf{EXH}}: \quad & q(a; e, f) = p(a|e, f; t) \\ \mathbf{M}^{\mathbf{MLE}}: \quad & t' = \arg\max_t Q(t, t^{old}) \\ & \quad = \arg\max_t \sum_{x,z} q(a|e, f) \log p(e, f, a; t) \end{aligned}$$

$$\mathbf{M}^{\mathbf{MAP}}: \quad t' = \arg\max_t Q(t, t^{old}) + \log p(t)$$

OOV words

- Prior Knowledge 3: Word alignment with prior knowledge about OOV words.
  - 1: Learn OOV word pairs by first detecting OOV words in source side after decoding parallel corpus, then by translating them by hands (transliteration, proper nouns, localization, symbols, equations and algorithms).
  - 2: Learn word alignment $P(e|f)$ in terms of the word pair frequencies (with alignment function as hidden variable) with the prior about an alignment function supplied by OOV word pairs..

$$align(e_i|f_i, T) = \begin{cases} 1 & (e_i = t_i, f_j = t_j) \\ 0 & (e_i = t_i, f_j \neq t_j) \\ 0 & (e_i \neq t_i, f_j = t_j) \\ \text{uniform} & (e_i \neq t_i, f_j \neq t_j) \end{cases}$$

Smoothing Methods for Small Corpus

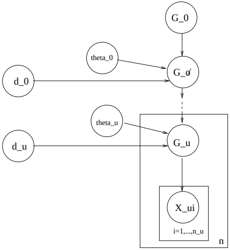
approach		reference	MT context
HPYLM	LM	[Goldwater&, 06]	[Okita&,10]
modified Kneser-Ney	LM	[Goodman&, 98]	
Good-Turing	LM		
Kneser-Ney	LM	[Kneser&,06]	
add-one	LM		
interporated Kneser-Ney	LM		
Bayesian smoothing	LM	[Peto&,05]	
HPYTM	TM		[Okita&,??]
Good-Turing	TM		[Foster&,06]
Kneser-Ney	TM		[Foster&,06]
others	TM		[Foster&,06]

Table: Smoothing Method for Language Model and Translation Model

Pitman-Yor Process Prior (1)

- Generative Model

$$
\begin{cases}
G_{\emptyset} | d_0, \theta_0, G_0 \sim PY(d_0, \theta_0, G_0) \\
\dots \\
G_u | d_{|u|}, \theta_{|u|}, G_{pi(u)} \sim PY(d_{|u|}, \theta_{|u|}, G_{pi(u)}) \\
X_i | G_u \sim DISCRETE(G_u), \qquad (i = 1, ..., n)
\end{cases}
$$



Pitman-Yor Process Prior (2)

- Pitman-Yor process as a prior:

$$G_u \sim PY(d_{|u|}, \theta_{|u|}, G_{pi(u)}) \quad (1)$$

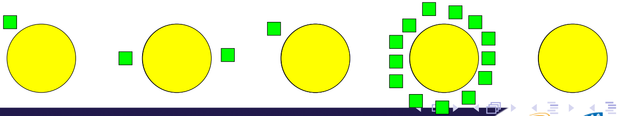
  - Pitman-Yor process $PY(d, \theta, G_0)$ is distribution over some base distribution G_0 [Pitman, 95], which has two parameters to generate a power-law distribution,
    - d: a discount parameter, and
    - θ: a strength parameter.
  - u: a given context.
  - $G_u(w)$: probability of the current word taking value w.
  - $pi(u)$: a function whose parameter is a context u,
  - the discount and strength parameters are functions of the length $|u|$ of the context,

Pitman-Yor Process Prior (3)

- Discount and strength parameters
  - First customer sits at the first available table.
  - ith subsequent customer sits at a table drawn from the following distribution (c_k: number of customers seated at table k until now, θ: this controls the similarity between G and G_0, and t: total number of tables until now):

$$
\begin{aligned}
P(\text{previously occupied table k}|\mathcal{F}_{i-1}) &\propto c_k - d \\
P(\text{the next unoccupied table}|\mathcal{F}_{i-1}) &\propto \theta + dt
\end{aligned}
$$

- Chinese restaurant (Customers enter and seat themselves):
  - Number of tables are infinite and
  - Each table has infinite seating capacity.



Pitman-Yor Process Prior (4)

- Let $pi(u)$ be the suffix of u consisting of all but the earliest word
  - u is n-gram words
  - $pi(u)$ is (n-1)-gram words
- Predictive distribution of n-gram probability in HPYLM
  - Recursively calculated as in (2):

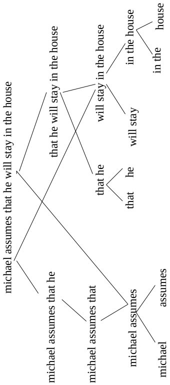
$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)} p(w|h') \tag{2}$$

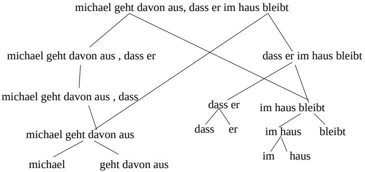
  - $p(w|h')$ is the same probability using a (n-1)-gram context h'.

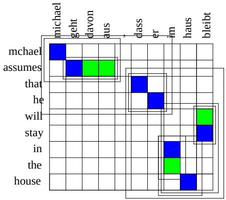
Pitman-Yor Process Prior (5)

- Incorporation of zero probabilities in decoder.
  - When we encounter unseen n-gram word in a test sentence, PB-SMT uses constant zero probabilities for unseen n-gram word,
  - HPYLM should look up different zero probabilities based on context, e.g. (n-1)-gram. $\longrightarrow$ Just before we do decoding, we update LM by supplying zero-probabilities for unseen n-gram word.

Pitman-Yor Translation Model







Pitman-Yor Translation Model

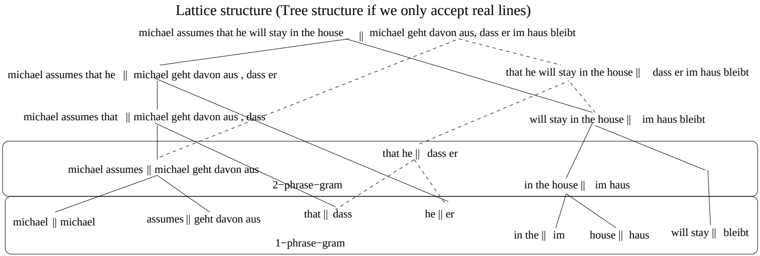


Figure: Figure shows a lattice structure of translation model for a toy example.

Statistics

JP–EN		num
prior knowledge 1	MWEs	120070
prior knowledge 2	paraphrases	432135
prior knowledge 3	transliteration	25928
	proper nouns	3408
	localisation	207
	equations	103
	symbols	13842
	noise	19007

n

Table: Statistics of prior knowledge.

Experimental Results

JP–EN	without TM smoothing	with TM smoothing
baseline	21.68	22.44
MWEs	22.48	22.78
paraphrases	22.23	22.44
OOVs	22.26	22.52
all	22.93	23.02
heuristics	21.90	22.49

Table: Results for 200k JP–EN sentences. Heuristics in the last row shows the result when prior knowledge 1 was added at the bottom of the translation model.

Table Of Contents

Factored Translation Model

- Factored translation model (Koehn and Huang, 2007; Avramidis and Koehn, 2008, Koehn, 2010)
    - factor: additional linguistic markup at the word level
        - morphological features including case, number, gender, person, tense, and aspect.

English Factors

Parsing results of "John has come" by Enju [Miyao and Tsujii, 2002]


```
<sentence id="s0" parse-status="success" fom="-0.0221762">
  <cons id="c0" cat="S" xcat="" head="c3" sem-head="c3" schema="subj-head">
    <cons id="c1" cat="NP" xcat="" head="c2" sem-head="c2" schema="empty-spec-head">
      <cons id="c2" cat="NX" xcat="" head="t0" sem-head="t0" >
        <tok id="t0" cat="N" pos="NN" base="john" lexentry="[D&lt;N.3sg&gt;]-lxm" pred="noun-arg0">
          John
        </tok>
      </cons>
    </cons>
    <cons id="c3" cat="VP" xcat="" head="c4" sem-head="c5" schema="head-comp">
      <cons id="c4" cat="VX" xcat="" head="t1" sem-head="t1">
        <tok id="t1" cat="V" pos="VBZ" base="have" lexentry="[NP&lt;V.have.bse&gt;VP.pap]-sctl-lxm-
singular3rd-verb-rule" pred="aux-arg12" aux="have" arg1="c1" arg2="c5">
          has
        </tok>
        <cons id="c5" cat="VP" xcat="" head="t2" sem-head="t2">
          <tok id="t2" cat="V" pos="VBN" base="come" lexentry="[NP.nom&lt;VP.bse&gt;]-lxm-
perfect-verb-rule" pred="verb-arg1" tense="present" aspect="perfect" voice="active" aux="minus" arg1="c1">
            come
          </tok>
        </cons>
      </cons>
    </cons>
  </cons>
</sentence>
```

Japanese Factors

Parsing results of “ジョンが来た”(John has come) by Cabocha [Kudo and Matsumoto, 2003]

0 1D 01 0.00000000						
ジョン	ジョン	ジョン	名詞-固有名詞-人名-名			B-PERSON
が	ガ	が	助詞-格助詞-一般			O
* 1 -1O 0/1 0.00000000						
来	キ	来る	動詞-自立	カ変・来ル	連用形	O
た	タ	た	助動詞	特殊・タ	基本形	O
EOS						

Centre for Next Generation Localisation

Word Alignment and Phrase Extraction

Word alignment in two directions

	null	ジョン	が	来	た
null			⇒		
John		⇓⇒			
has			⇓		
come				⇓⇒	

↓

Phrase extraction

John	has	come	
ジョン	が	来	た

Factored Translation Model

Translation process (Lemma)

	ジョン	が	来る	た
John	⇓⇒			
have		⇓		
come			⇓⇒	

Translation process (POS)

	名詞	助詞	動詞	助動詞
noun	⇓⇒			
aux		⇓		
verb			⇓⇒	

Translation process (Morphology)

	無変化	無変化	連用形	基本形
singular	⇓⇒			
singular3rd		⇓		
present perfect			⇓⇒	

Generation process

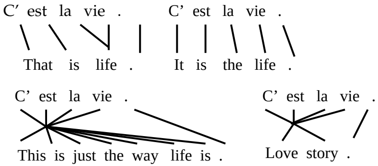
	ジョン	が	来	た
ジョン,名詞,–	↗			
が,助詞,無変化		↗		
来る,動詞,連用形			↗	
た,助動詞,基本形				↗

Factored Translation Model

Problematic cases (n-to-m mapping problems including paraphrases, MWEs, non-literal translations)

	c’	est	la	vie
that				
is				
just	$\Downarrow\Rightarrow$?			
the				
way				
life				$\Downarrow\Rightarrow$?
is	$\Downarrow\Rightarrow$?			

	c’	est	la	vie
love				$\Downarrow\Rightarrow$?
story	$\Downarrow\Rightarrow$?			



Strong Assumptions

1. Correct word correspondence. (Precision $\leq$ 90%)
2. Decision to separate morpheme(s) from a word. (e.g. Separate 'looks' into 'look' and 's')
   - Separation of word(s) and morpheme(s): better BLEU score, (adequacy decreases).
   - Combination of word(s) with morpheme(s)
3. Sufficient morphological information for (monolingual) language.
   - Morphological info (most of the verbs in European language inflect based on person and number, Japanese verbs inflect based on aspect).
   - Some missing morphological info (No article and gender for noun phrases in Japanese.)

Japanese

- Inflection in verbs / adjectives / adverbs
  - Conjugation in six stem forms (imperfective / continuative / terminal / attributive / hypothetical / imperative form) based on aspect.
- JP noun phrases are accompanied with case particles.
- (Relatively) free word order
- SOV language

Algorithm to Change Assumptions

Overall Algorithm

1. (Morphological predesign)
2. Do segmentation of JP sentence into morphemes by a morphological analyzer.
3. Combine verb and morphemes with attaching case information. By this construction, we aim at not losing the information by morphemes).
4. Do word alignment by MWE-sensitive word aligner (Okita et al., 2010) instead of GIZA++.
   - Plugged in prior knowledge about bilingual terminology (Bilingual terminology extraction algorithm)

Experimental Settings (1)

- Baseline: Log-linear PB-SMT
  - GIZA++ [Och and Ney, 2003] of IBM Model 4.
  - Phrase extraction: grow-diag-final heuristics.
  - SRILM [Stolcke, 2002]: 5-gram language model.
  - MERT [Och, 2003].
  - Moses decoder [Koehn et al., 2007].
- Corpus
  - NTCIR-8 corpus [Fujii, 2010]. EN–JP. Training corpus 200k (randomly extracted); development set 1,200 sent; test set 1,119 (EN–JP) / 1,251 (JP–EN).

Experimental Results

- Baseline by the plain factored model: 21.67 BLEU point absolute.
- Separation of a word with morpheme(s) : 18.35 BLEU point absolute.
- MWE-sensitive word aligner: 22.23 BLEU point absolute.

Conclusions and Further Works

- MWE-sensitive word aligner worked for factored translation model as well.
- Need more translation pairs and bigger size.
  - (MWE-sensitive word aligner) Searched alignment space should be enlarged.
  - Need some strategy to handle free word order.

Acknowledgement

Thank you for your attention.

- This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.
- Irish Centre for High-End Computing.
- Machine Translation Marathon.

MWE-sensitive Word Alignment (7)

Algorithm 1 MWE Extraction Algorithm (Step 1 in overall alg)

Given: a parallel corpus and a set of anchor word alignment links:

1. We use a POS tagger to tag a sentence on the SL side.
2. Based on the typical POS patterns for the SL, extract noun phrases on the SL side.
3. Count n-gram statistics (typically $n = 1, \cdots, 5$) on the TL side which jointly occur with each source noun phrase extracted in Step 2.
4. Obtain the maximum likelihood counts of joint phrases, i.e. noun phrases on the SL side and n-gram phrases on the TL side.
5. Repeat Steps 1 to 4 reversing SL and TL.
6. Intersect (or union) the results in both directions.

Statistics

observed	#	%	type	#
1 form	911	40%	NP	1831012
2 forms	445	20%	VP	259432
3 forms	506	22%	ph (symbols)	68298
4 forms	270	12%	ph (prefixes)	66729
5 forms	111	5%	ph (OOVs)	66461
6 forms	33	1%	ph (conjunctions)	65159
			ph (attributives)	59633
			Adverbial phrases	33781

Table: Statistics of observed verb forms (left) and number of phrase types(right) in JP side. In right figures, the inside of parenthesis means that the top of the phrase starts with symbols, and so forth.

Statistics

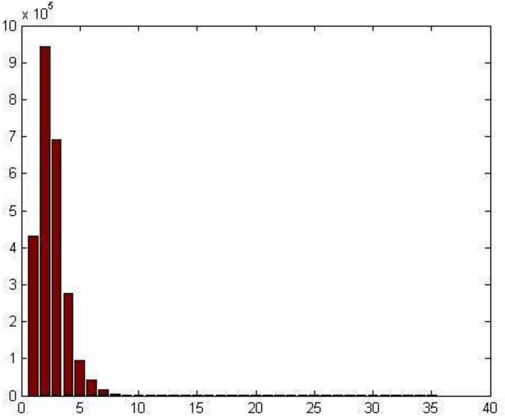


Figure: Statistics of number of nouns in NP.