Very Large-scale Lexica (VLSL) to Enhance Entity Translation in MT Systems

TAUS Executive Forum Tokyo 2017

Jack Halpern
CEO, The CJK Dictionary Institute, Niiza, Japan

Overview

- recognition and translation of named entities is major issue
- state-of-the-are technology inadequate for entity translation
- high accuracy NET/NER requires large scale lexica
- large scale resources expensive to build and maintain
- robust MT systems need need large-scale computational lexica

Very Large Scale Lexical Resources (VLSLR)

- millions of CJK named entities
- multilingual database of Japanese POIs and place names
- comprehensive multilingual database of Chinese personal names
- CJK and Arabic lexical databases

High Failure Rates for Japanese POIs

- Even state-of-the-art NMT and NER systems often fail to recognize and accurately process entities
- Spot tests conducted on Japanese POIs using major MT translation services
- failure rates are 55% for Google, 62% for Bing, and 48% for Baidu

Japanese	Google NMT	Bing Translator	Baidu Translate	Correct translation by CJKI
海の中道線	Midair line of the sea	The middle line of the sea	The sea line	Umi-no-Nakamichi Line
三角線	Triangle	Triangular line	Misumi	Misumi Line
神津島空港	Kozu Island airport	God Tsushima Airport	Kozu Island Airport	Kozushima Airport
孔子公園	Confucius Park	Confucius Park	Confucius Park	Koshi Park
手取フィッシュランド	Takeshi Fishland	Fish Land	Tedori fish Landes	Tedori-fishland
パレマルシェ神宮	Palermark Shinto shrine	Palais Marche Jingu	Palais du Marche Shrine	Pare Marché Jingu

Comprehensive Database of Japanese POIs and Place Names

- CJK, European, and other Asian languages like Thai, Vietnamese, and Indonesian
- approximately 3.1 million entries spread over 14 languages
- numerous POI types, including schools, highways, train stations and commercial facilities

Comprehensive Database of Japanese POIs and Place Names

日中韓英阿 (CJKEA)					
Japanese	**Chinese (SC)**	**Chinese (TC)**	**Korean**	**English**	**Arabic**
成田国際空港	成田国际机场	成田國際機場	나리타국제공항	Narita International Airport	مطار ناريتا الدولي
那覇空港	那霸机场	那霸機場	나하공항	Naha Airport	مطار ناها
東京駅	东京站	東京站	도쿄역	Tokyo Station	محطة طوكيو
京都府庁	京都府厅	京都府廳	교토부청	Kyoto Prefectural Office	مكتب محافظة كيوطو
大阪市	大阪市	大阪市	오사카시	Osaka City	مدينة أوساكا

Comprehensive Database of Japanese POIs and Place Names

アジア言語等 (Asian)					
Japanese	**Indonesian**	**Vietnamese**	**Thai**	**Hindi**	**Russian**
成田国際空港	Bandar Udara Internasional Narita	Sân bay quốc tế Narita	ท่าอากาศยานนานาชาตินะริตะ	नारिता अं तर्रा ष्ट्रीय हवाई अड्डा	Международный аэропорт Нарита
那覇空港	Bandar Udara Naha	Sân bay Naha	ท่าอากาศยานนะฮะ	नाहा हवाई अड्डा	аэропорт Наха
東京駅	Stasiun Tokyo	Ga Tokyo	สถานีโตเกียว	टोक्यो स्टे शन	станция Токио
京都府庁	Kantor Pemerintahan Kyoto	Tòa nhà chính quyền tỉnh Kyoto	ที่ว่าการนครเกียวโต	क्योतो प्रीफ़े क्चर मु ख्यालय	администрация префектуры Киото
大阪市	Kota Osaka	Thành phố Osaka	อำเภอโอซากา	ओसाका सिटी	город Осака

Comprehensive Database of Japanese POIs and Place Names

ヨーロッパ言語 (European)					
Japanese	**German**	**Portuguese**	**Spanish**	**French**	**Italian**
成田国際空港	Internationaler Flughafen Narita	Aeroporto Internacional de Narita	Aeropuerto Internacional de Narita	Aéroport international de Narita	Aeroporto Internazionale di Narita
那覇空港	Flughafen Naha	Aeroporto de Naha	Aeropuerto de Naha	Aéroport de Naha	Aeroporto di Naha
東京駅	Bahnhof Tokio	Estação de Tóquio	Estación de Tokio	Gare de Tokyo	Stazione di Tokyo
京都府庁	Präfekturverwaltung Kyoto	Sede do Governo de Quioto	Oficina Prefectural de Kyoto	Préfecture de Kyoto	Sede del Governo prefettizio di Kyoto
大阪市	Stadt Osaka	Cidade de Osaka	Ciudad de Osaka	Ville d'Osaka	Città di Osaka

Very-Large Scale CJK and Arabic Lexical Databases

- monolingual word databases for both Simplified and Traditional Chinese, Japanese, Korean
- separate word databases for canonical and full-form Arabic
- covers general vocabulary, proper nouns and technical terms
- databases cover in total roughly 27,500,000 entries

Very-Large Scale CJK and Arabic Lexical Databases

Language	General vocabulary	Proper nouns	Technical terms	Total
Arabic	113,973	95,184	0	209,157
Arabic (full form)	14,452,336	95,184	0	14,547,520
Japanese	459,980	1,017,221	1,169,652	2,646,853
Korean	83,835	42,280	914,772	1,040,887
Simplified Chinese	1,395,979	1,730,881	2,153,157	5,280,017
Traditional Chinese	1,731,030	48,527	2,153,157	3,932,714
Total	**18,237,133**	**3,029,277**	**6,390,738**	**27,657,148**

Arabic Lexical Database – Canonical Forms

Type	POS	Unvocalized Arabic	Vocalized Arabic	Phonemic Transcription
G	V	قحط	قَحَّط	qáḥḥaṭ
G	V	قطط	قَطَّط	qáṭṭaṭ
G	N	قفر	قَفْر	qafr
G	N	قمل	قَمْل	qaml
G	N	قريحة	قَرِيحَة	qarīḥa
G	V	قيل	قَيَّل	qáyyal
G	N	قرى	قِرَى	qíraˌ
G	N	قصور	قُصُور	quṣūr
G	N	قرعة	قُرْعَة	qúrɛa
G	N	رذيلة	رَذِيلَة	radhīla

Arabic Lexical Databases - Full Forms

Type	POS	Unvocalized Arabic	Vocalized Arabic	Phonemic Transcription
G	V	أكتب	أَكْتُبَ	ʾáktuba
G	V	أكتب	أَكْتُبْ	ʾáktub
G	V	أكتب	أَكْتُبُ	ʾáktubu
G	V	أكتب	أُكْتَبَ	ʾúktaba
G	V	أكتب	أُكْتَبْ	ʾúktab
G	V	أكتب	أُكْتَبُ	ʾúktabu
G	V	اكتبا	اُكْتُبَا	ʾúktubạ
G	V	اكتب	اُكْتُبْ	ʾúktub
G	V	اكتبن	اُكْتُبْنَ	ʾuktúbna
G	V	اكتبوا	اُكْتُبُوا	ʾúktubụ

Japanese Lexical Database

Type	POS	Headword	Reading
G	NC	啓知児	ケイチジ
G	VN	啓発	ケイハツ
G	NC	啓蒙哲学	ケイモウテツガク
G	NC	啓塾	ケイチツ
G	NC	珪華	ケイカ
G	NC	珪涌	フンヨウ
G	V1	型にはめる	カタニハメル
G	NC	型押し文	カタオシモン
G	NC	型構造	カタコウゾウ
G	NC	型持ち	カタモチ

Korean Lexical Database

Type	Headword
G	응축
G	응전하다
G	응고되다
G	응모
G	응봉
G	응답하다
G	응용되다
G	읍촌
G	와지끈뚝닥
G	와해하다

Simplified Chinese Lexical Database

Type	POS	Headword	Pinyin
G	N	天帝	tian1-di4
G	N	天道	tian1-dao4
G	N	天南星	tian1-nan2-xing1
G	N	天年	tian1-nian2
G	U	天秤	tian1-cheng4
G	E	天不盖，地不载	tian1-bu4-gai4-di4-bu4-zai4
G	N	天文	tian1-wen2
G	N	天文馆	tian1-wen2-guan3
G	U	天文照相望远镜	tian1-wen2-zhao4-xiang4-wang4-yuan3-jing4
G	N	天文物理学	tian1-wen2-wu4-li3-xue2

Traditional Chinese Lexical Database

Type	POS	Headword	Pinyin
G	N	標兵	biao1-bing1
G	N	標本館	biao1-ben3-guan3
G	N	標本商	biao1-ben3-shang1
G	N	標名	biao1-ming2
G	N	標目	biao1-mu4
G	V	標會	biao1-hui4
G	N	標價卡	biao1-jia4-ka3
G	N	標售制	biao1-shou4-zhi4
G	N	標售物	biao1-shou4-wu4
G	N	標幟	biao1-zhi4

Okurigana Variants

Headword	Reading	Normalized
書き著す	かきあらわす	書き著す
書き著わす	かきあらわす	書き著す
書著す	かきあらわす	書き著す
書著わす	かきあらわす	書き著す

Variants of *toriatsukai*

toriatsukai	Type of Variant
取り扱い	"standard" form
取扱い	okurigana variant
取扱	All kanji
とり扱い	replace kanji with hiragana
取りあつかい	replace kanji with hiragana
とりあつかい	All hiragana

Strategies for Japanese-English POI Conversion

Example:「東京中央ゴルフ場」			
	翻訳レベル	**翻訳例**	**解説**
1	音訳1	Tokyo Chuo Gorufujo	完全な音訳。読みさえあれば、ほぼ自動処理が可能。
2	音訳2	Tokyo Chuo Golf-jo	音訳＋外来語由来のカタカナのみ正しいスペルを調査。語源不明の造語等は音訳で処理する場合もある。
3	意音訳	Tokyo Chuo Golf Course	ゴルフ場 =″Golf Course″、病院 =″Hospital″、デパート =″Department Store″ のように大まかな訳語を定めておき、それ以外の部分を音訳する。
4	意訳	Tokyo Central Golf Course	固有名以外の部分を全て意訳する。
5	翻訳	The Central Golf Club, Tokyo	ウェブサイトや電話調査で、個々の正式訳を調査する。現地での表示と一致するので最も望ましい形だが、作成の時間・経費が嵩む。

Strategies for Japanese-Chinese POI Conversion

Example: 「幕張国際展示場」				
Type	**English**	**Simplified Chinese**	**Traditional Chinese**	**Comments**
字訳	---	幕张国际展示场	幕張國際展示場	日本語表記を文字対応レベルで変換
音訳	Makuhari kokusai tenji jo	–––	- - -	日本語読みの音通りに表記
意訳	Makuhari International Exhibition Area	幕张国际展览馆	幕張國際展覽館	地名等の固有名以外、全て意訳
意音訳	Makuhari Kokusai Exhibition Area	–––	- - -	主要な構成要素部分を意訳、それ以外は音訳
翻訳	International Exhibition Halls, Makuhari Messe	幕张国际展览中心	幕張國際展覽中心	施設自体が公式に定めた名称

Cross-Script Variants

Type	Example
Kanji vs. Hiragana	大勢　おおぜい
Kanji vs. Katakana	硫黄　イオウ
Kanji vs. hiragana vs. katakana	猫　ねこ　ネコ
Katakana vs. hybrid	ワイシャツ　Ｙシャツ
Kanji vs. katakana vs. hybrid	皮膚　ヒフ　皮フ
Kanji vs. hybrid	彗星　すい星
Hiragana vs. katakana	ぴかぴか　ピカピカ

The Three Conversion Levels

Level 1	Code	Character-to-character, code-based substitution
Level 2	Orthographic	Word-to-word, character-based conversion
Level 3	Lexemic	Word-to-word, lexicon-based conversion

Code Conversion

SC	TC1	TC2	TC3	TC4	Remarks
门	門				one-to-one
汤	湯				one-to-one
发	發	髮			one-to-many
暗	暗	闇			one-to-many
干	幹	乾	干	榦	one-to-many

Orthographic Conversion

English	SC	TC1	TC2	Incorrect	Comments
telephone	电话	電話			unambiguous
we	我们	我們			unambiguous
start-off	出发	出發		出髮 齣髮 齣發	one-to-many
dry	干燥	乾燥		干燥 幹燥 榦燥	one-to-many
	阴干	陰乾	陰干		depends on context

Lexemic Conversion

English	SC	Taiwan TC	Hong Kong TC	Incorrect TC
Software	软件	軟體	軟件	軟件
File	文件	檔案	檔案	文件
Program	程序	程式	程式	程序
Taxi	出租汽车	計程車	的士	出租汽車
Osama Bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹	奧薩馬本拉登
Kennedy	肯尼迪	甘迺迪	堅尼地	肯尼迪

Database of 100 Million Chinese Personal Names

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
	Simplified Chinese	艾	单	业经	爱博
Simplified Chinese	Toned Pinyin	ài	shàn	yèjīng	àibó
	Numbered Pinyin	ai4	shan4	ye4-jing1	ai4-bo2
	Wade-Giles	ai	shan	yehching	aipo
	Yale System	ai	shan	yejing	aibwo
	Tongyong	ai	shan	yejing	aibo
Traditional Chinese	Traditional Chinese	艾	單	業經	愛博
	Zhuyin	ㄞˋ	ㄕㄢˋ	ㄧㄝˋ ㄐㄧㄥ	ㄞˋㄅㄛˊ

Database of 100 Million Chinese Personal Names

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
Simplified Chinese		艾	单	业经	爱博
Cantonese	LAU1	ngaai6	daan1	yip6-ging1	ngoi3-bok3
	LAU2	ngaai	daan	yipging	ngoibok
	YALE1	ngaai6	daan1	yip6-ging1	ngoi3-bok3
	YALE2	ngaai	daan	yipging	ngoibok
	JYUTPING1	ngaai6	daan1	jip6-ging1	ngoi3-bok3
	JYUTPING2	ngaai	daan	jipging	ngoibok
Hokkien		gai	sean	yapkeng	ai po
Hakka		ngioi	shan	ngiapgin	oibok

Database of 100 Million Chinese Personal Names

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
Japanese	Japanese	艾	単	業経	愛博
	Hiragana	あい	だん	ぎょうけい	あいはく
	Katakana	アイ	ダン	ギョウケイ	アイハク
Korean	Korean Hanja	艾	單	業經	愛博
	Hanja reading	아이	산	예징	아이보
	Korean Hangul	애	단	업경	애박
	MOE	ai	san	yejing	aibo
	NRS	ai	san	yejing	aibo
	KLS	ai	san	yejing	aibo
	ISO DPRK	ai	san	yecing	aipo
	ISO ROK	ai	san	yejing	aibo
Other languages	Vietnamese	ngải / nghệ	đơn / đan	Nghiệp Kinh	ái bác
	English	ai	shan	yejing	aibo

Thank You

شكرا جزيلا

谢谢

ありがとうございました

감사합니다