Lexicographic Criteria for Selecting Multiword Units for MT Lexicons

presented by
Jack Halpern, CEO
The CJK Dictionary Institute
日中韓辭典研究所

MUMTTT 2019

The 4th Workshop on Multi-word Units in Machine Translation and Translation Technology

September 27, 2019, Malaga

Overview

- The fundamental principles for identifying and selecting MWUs
- Defining the various subtypes of MWU based on lexicographic principles
- Large-scale resources can significantly enhance the translation accuracy of multiword proper nouns

Multiword Unit Subtypes

1. **Multiword expression (MWE)**: a lexical unit consisting of two or more words that together function as a single lexical unit.
2. **Free word combination (FWC)**: a meaningful free sequence of words that follow the rules of syntax but has no lexical status.
3. **Phraset**: a recurrent meaningful free combination of words that has no lexical status in the source language but corresponds to a lexical unit in the target language.
4. **Collocation**: a recurrent combination of words co-occurring more often than by chance whose meaning is (mostly) compositional and transparent.
5. **Multiword proper noun**: a combination of two or more words that together function as a single proper noun.

Multiword Expressions (MWEs)

zona residencial	residential zone (transparent compositional compound)
dar a	look out onto (opaque non-compositional phrasal verb)
elefante blanco	white elephant (opaque bilingually compositional idiom)
devanarse los sesos	rack one's brains over (idiomatic expression)
matar dos pájaros de un tiro	kill two birds with one stone (opaque compositional proverb)
lo antes posible	as soon as possible (locution)

Free Word Combination (FWC)

A free word combination (FWC) is a meaningful free sequence of words that follow the rules of syntax but has no lexical status.

drink water
cerrar con las manos
abrir un agujero
abrir la luz
write a poem
don't come home

FWCs vs. MWEs

FWCs	
abrir un agujero	dig a hole
abrir un túnel	dig a tunnel
abrir la luz	turn on the light
abrir el agua	turn on the water

MWEs	
abrir la puerta	open the door
abrir un hospital	open a hospital
abrir el baile	begin the dance

Phraset

A phraset is a free, meaningful combination of words (FWC) that is recurrently used to express a concept that has no lexical status but corresponds to a lexical unit in another language.

Examples	
cerrar con llave	to lock
ir en bicicleta	to cycle
ir en coche	to go by car
ir en monociclo	to unicycle
ir en avión	to fly

Collocation

A collocation (or institutionalized phrase) is a recurrent combination of words that co-occur more often than by chance whose meaning is (mostly) compositional and transparent.

Examples	
bonita sorpresa	nice surprise
estar fascinado con	be fascinated with
tomar una decisión	make a decision
hacer una pausa	take a break
prestar atención	pay attention
hacer amor a/com	make love to/with
respecto a	with regard to
abandonarse a la desperación	to fall into despair

Multiword Proper Nouns

A multiword proper noun is a combination of two or more words that together function as a single proper noun.

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**

Inaccurate Translations of POIs

Recognition and accurate translation of proper nouns, many of which are bilingually non-compositional, are a major issue in MT and other NLP applications.

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

Thank You
ありがとうございます
Muchas gracias
Muito obrigado

The CJK Dictionary Institute
Niiza-shi, Saitama, Japan

www.cjk.org