Can AI do Term Extraction as well as Humans?

Compiled and presented by Steen Kesmodel

AI prompting by Inacio Vieira and Lana Taratukhina

Alpha CRC

memoqfest

memoqfest

Term Extraction?

memoqfest

Traditional Term Extraction

Two methods

- Use Extraction tool to generate list of *candidates*
  - Translator/Terminologist to select from list
  - Mostly based on *frequency*
- Read through corpus and extract as you go

- Both methods are labour intensive
  - Angelica Zerfaß: 2 days for 20k corpus

General rules for Term Extraction

The Journal of Specialised Translation Issue 18 – July 2012

Terminology in the age of multilingual corpora
Alan K. Melby, Brigham Young University[1]

ABSTRACT

Terminology management has long played an important role in translation and localisation. It has been asserted, however, that the need for terminology management is declining with the rise of widely accessible aligned multilingual corpora, such as bi-texts. In this view, translators will be able to identify terms and their translations by using previous translations to automatically identify the best translation for a term. This article, however, argues that while bi-text resources will assist in human-oriented terminology management, they will actually increase the need for skilled terminology work and termbases. Furthermore, because more tools will generate terminological data, the need for exchange between tools will increase. After discussing the case for terminology management and terminology exchange in the age of aligned multilingual corpora, the paper describes the role of the TermBase eXchange (TBX) standard in terminology exchange, including typical scenarios for its use, and some of the challenges faced in using it.

KEYWORDS

Concept-oriented terminology, data exchange, glossary, interoperability, aligned multilingual corpus, multilingual corpus, bi-text, TBX, termbase, TermBase eXchange, terminology, terminology management.

"There is unfortunately no cure for terminology; you can only hope to manage it."
— Kelly Washbourne (personal communication)

1. Introduction

In recent years corpus linguistics has seen tremendous success in translation-related fields. For the purposes of this article the term

1

Good summary in

Terminology in the age of multilingual corpora
By Alan K. Melby, Brigham Young University, July 2012

2

Rules for *terms* e.g.

Domain and/or Company specific, mostly nouns and phrases, avoid common words, consider *frequency*, stop-words, etc.

3

Where there are rules, there are *interpretations*!

memoqfest

The Experiment

1. Take a corpus: 983 UI strings 5325 words
2. Have a Human do Term Extraction
3. Ask some AIs to do the same
4. Compare results
5. Conclusion...

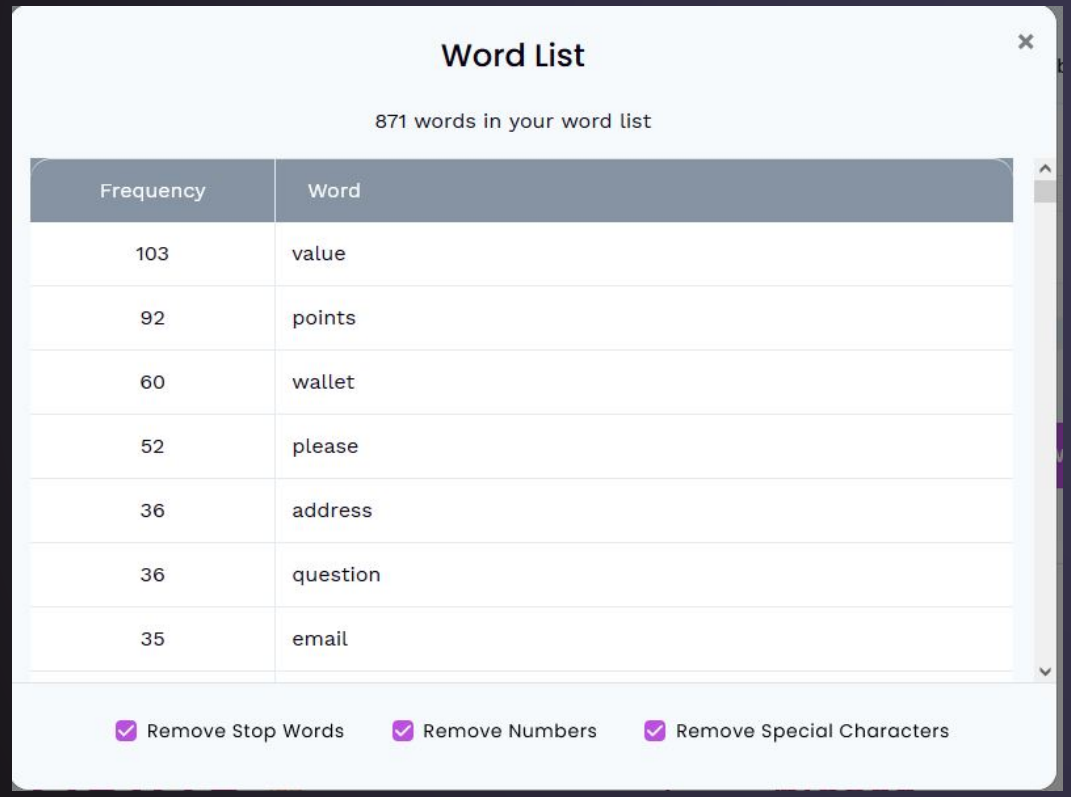
memoqfest

The Experiment: Human I

ID	Context	English
100040	textNicknameRegular	A nickname is 4-30 characters long without spaces.
100041	textNicknameDescription	A nickname should be 4-30 characters long, and may contain Chinese characters, English letters, numbers, underscores and minus signs.
100042	textPasswordRegular	A password should be 6-14 characters long and include at least a combination of letters, numbers and symbols.
100043	textValidTimeDescription	A question will be closed upon expiration if it does not become public; the closing time for public questions will be set uniformly as per the platform rules.
100044	textSeedPhraseDescription	A seed phrase consists of words separated by spaces.
100045	btnAccept	Accept
100046	error10001	Account already exists
100047	textAccountBalance	Account balance
100048	textAccountBindSuccess	Account bound successfully
100049	textAccountDeleted	Account deactivated
100050	error10005	Account does not exist
100051	error10051	Account has been deleted
100052	error10002	Account is frozen
100053	btnAdd	Add
100054	btnAddReason1	Add
100055	textSetupAddEmailDescription	Add an email address used to log in, receive notifications and reset passwords.
100056	btnAddEmailAddress	Add email address
100057	textAddEmailAddress	Add email address
100058	textSetupAddEmail	Add email address
100059	textAddLink	Add link
100060	btnAddOption	Add option
100061	btnAddReason	Add reason
100062	textAddReason	Add reason
100063	textAddress	Address
100064	textAdvancedBox	Advanced Mystery Box
100065	textLikesDescription	After a question ends, the top 5 reasons with the most likes will win, and those who gave likes to the winning reasons can share the points.
100066	btnAlbum	Album
100067	btnAll	All
100068	tagAll	All
100069	textAllChannel	All channels
100070	textFilterReason	All reasons
100071	textShownAll	All shown

- Using the ‘read through and extract’ method
- A translator did this as part of a job for a client
- I received the list….
- Some choices made me wonder
- So I did it myself for reference…

The Experiment: More Humans!



- Having now 2 Humans revealed the inexact nature of term Extraction:
- One extracted 73 terms, the other 17
- 9 terms in common
- So which one is more *correct/best/most useful?*
- Let 3 more translators do the extraction…

memoqfest

The Experiment: Five Humans I

Human 1 (H1)	Human 2 (H2)	Human 3 (H3)	Human 4 (H4)	Human 5 (H5)	
59%	76%	59%	77%	50%	Strike rate: #common/Total selected
62%	19%	59%	43%	81%	Success rate: #common/all common
73	17	69	39	112	Total # terms selected

- Total number of terms selected by each (yellow, 311 different terms in total)
- Which criteria for inclusion in *baseline*?
- Terms selected by 2 or more (literally the *lowest common denominator*)
- Consensus is a rare thing!

Picked by 2 or more	Picked by 3 or more	Picked by 4 or more	Picked by all 5
69	32	12	3
22% of all selected terms	16%	6%	1.5%

memoqfest

The Experiment: Five Humans I

Human 1 (H1)	Human 2 (H2)	Human 3 (H3)	Human 4 (H4)	Human 5 (H5)	
59%	76%	59%	77%	50%	Strike rate: #common/Total selected
62%	19%	59%	43%	81%	Success rate: #common/all common
73	17	69	39	112	Total # terms selected

Strike rate: the percentage of terms selected by **two or more/Total number** of selected words by this translator (an indication of *efficiency*)
Success rate: the percentage of terms also **selected by someone else/all selected by two or more** (a measure of how big a proportion their picks are of the total list, we might call it *accuracy*)
Total number of terms selected: actual number of terms picked by each translator

memoqfest

The Experiment: Humans round II

- What if instead of picking from the corpus, they were to pick from the list of all terms selected by anyone in round one? (Like selecting from a list of 311 *candidates*)
- The complete new list has 195 different entries

Picked by 2 or more (II)	Picked by 3 or more	Picked by 4 or more	Picked by all 5
187	144	82	23
95% of all selected terms	73%	42%	12%

Picked by 2 or more (I)	Picked by 3 or more	Picked by 4 or more	Picked by all 5
69	32	12	3
35% of all selected terms	16%	6%	1.5%

- More terms are picked by everyone
- Both in absolute and relative numbers

The Experiment: AI

- Selected 4 AIs
- Alpha's own (ChatGPT based), Termxt (2x) Claude (3x) ChatGPT (2x)
- Created prompts, based on Melby's article
- Some then prompted differently (e.g. number of terms expected)

memoqfest

The Experiment: AI

Alpha CRC Prompt	TERMXT - 150 terms	TERMXT - 80 terms	Claude prompt 1	Claude prompt 2	Claude 3 - UI focussed	ChatGPT	ChatGPT		Human Common 2-5	
9%	9%	9%	1%	15%	9%	16%	15%			strikerate: #common/Total selected
14%	20%	10%	1%	22%	19%	13%	16%			success rate: #common/all common
106	150	80	87	101	139	57	75			Total terms extracted
10	14	7	1	15	13	9	11		69	Total # terms common w Humans

- Total number of terms selected by each (yellow)
- Below number of terms in common with Human extraction (2+)

Strike rate: measure of **efficiency**
Success rate: measure of **accuracy**

Picked by 2 or more AIs	Picked by 3 or more	Picked by 4 or more	Picked by all 5
214	32	12	3
26% of all selected terms (802)	4%	1,5%	0,4%

Picked by 2 or more H I	Picked by 3 or more	Picked by 4 or more	Picked by all 5
69	32	12	3
22% of all selected terms	16%	6%	1.5%

Picked by 2 or more (II)	Picked by 3 or more	Picked by 4 or more	Picked by all 5
187	144	82	23
95% of all selected terms	73%	42%	12%

memoqfest

The Experiment: AI part II 6 months later

- Improvements in AI led us to try another round with the AI
- Gemini AI Pro (Google), Claude Sonnet (Anthropics), Claude Opus (Anthropics), Llama-3 - 70B (Meta AI), GPT4-Turbo (Open AI)

Picked by 2 or more AIs I	Picked by 3 or more	Picked by 4 or more	Picked by all 5
214	32	12	3
26% of all selected terms (802)	4%	1,5%	0,4%

Picked by 2 or more AIs II	Picked by 3 or more	Picked by 4 or more	Picked by all 5
88	26	13	4
19% of all selected terms (452)	3%	3%	1%

memoqfest

The Experiment: AI vs Humans

How did the numbers stack up?

Terms in common with baseline	Human I 2-5	Human II 2-5	AI 2-8	AI 2024 2-5
Human I 2-5 (baseline 69)		69	40	44
Human II 2-5 (baseline 179)	69		57	40
Efficiency				
Strike rate: #common with baseline/Total common by group	Human I 2-5	Human II 2-5	AI 2-8	AI 2024 2-5
Human I 2-5 (baseline)		39%	19%	50%
Human II 2-5 (baseline)	39%		27%	45%
Accuracy				
Success rate: #common with baseline/baseline	Human I 2-5	Human II 2-5	AI 2-8	AI 2024 2-5
Human I 2-5 (baseline)		100%	58%	64%
Human II 2-5 (baseline)	39%		30%	21%

memoqfest

The Experiment: AI vs Humans

How did the numbers compare?

Strike rate: #common with baseline/Total common by group	Best Human	Worst Human	Best AI	Worst AI
Human I 2-5 (baseline)	77%	50%	16%	1%
Human II 2-5 (baseline)	93%	93%	54%	29%
Success rate: #common with baseline/baseline	Best Human	Worst Human	Best AI	Worst AI
Human I 2-5 (baseline)	81%	19%	22%	1%
Human II 2-5 (baseline)	100%	22%	11%	25%

Conclusion?

1. Humans more consistent when picking from a list made by humans
2. Very little consensus in interpreting which terms should be extracted
3. AI "*worse*" at finding common terms
4. AI no worse than most individual humans
5. AI has very good strike and success rates in some cases
6. Quality vs Quantity: which words and phrases were left out/included?

memoqfest

Appendix: The Term Extraction Guidelines

The guidelines for both translators and later AI, were based paper on glossary creation by linguist Alan K. Melby, titled "Terminology in the Age of Multilingual Corpora".

- Focus on domain-specific terminology - Extract words that pertain specifically to the subject matter or field that the text covers. Generic words used across domains are less relevant.
- Prioritize nouns over verbs and adjectives - Nouns tend to be the core terminology that requires consistent translation. Verbs and adjectives may vary more across languages.
- Consider multi-word terms as well as single words - Technical terminology often consists of multi-word noun phrases that should be treated as a unit.
- Note terms that may have multiple meanings - Homonyms that have a different meaning within the domain versus in general language should be flagged.
- Watch for inconsistent use of synonyms - If the original author uses different terms for the same concept, this should be documented.
- Exclude generic function words - Words like articles, prepositions, pronouns, etc. can be ignored as they likely have standard translations.
- Record acronyms and abbreviations - Any abbreviated forms need clear explanations for accurate translation.
- Check for spelling variations - Spelling errors and alternate spellings affect term extraction and should be fixed or noted.
- Consider hierarchical relationships - Broader, narrower and related terms may need translation as a group.
- Note any customer/project specific terms - Words unique to a particular usage context are important to flag.
- The goal is to systematically extract the vital terminology from the source text in a format that helps human translators stay consistent.

memoqfest

Thank you!

Any questions?

skesmodel@alphacrc.com