PARALLEL ANNOTATED SYNTHETIC CORPORA (PASC)

注釈付き対訳合成コーパス

Jack Halpern (jack@cjki.org)
The CJK Dictionary Institute, Inc.

Abstract

This paper introduces the Parallel Annotated Synthetic Corpora (PASC) project, an innovative approach to bolstering natural language processing (NLP) applications. Developed by The CJK Dictionary Institute, PASC leverages supervised generation techniques to create multilingual synthetic corpora with meticulous linguistic precision. PASC stands out for its fully aligned bilingual/multilingual content, accurate transcriptions, and rich annotations. These attributes enhance translation accuracy and support applications such as neural machine translation (NMT), automatic speech recognition (ASR), and text-to-speech (TTS). By addressing resource scarcity, PASC presents a promising avenue for advancing NLP technologies.

1. Introduction

1.1 Artificial corpora

Data augmentation is the process of adding to a dataset modified copies of existing elements, or creating new data based on existing data. Data augmentation can be performed on corpora, in which case they are referred to as **augmented corpora** (AC). These have become a hot topic in the machine translation community.

On the other hand, *synthetic data* is artificial data that mimics real-world observations and is used to train machine learning models when actual data is difficult or expensive to obtain. Corpora created artificially from synthetic data are referred to as **synthetic corpora** (SC) or artificial corpora.

1.2 Benefits to NLP and MT

There are various natural language processing (NLP) applications that can benefit from the augmentation or synthetization of corpora for use in language model training, such as neural machine translation (NMT), automatic speech recognition (ASR), text to speech (TTS), and others.

This can be exemplified by looking at Hasan et al., who have reported significant improvements, 2.8 points on the BLEU scale, in their NMT-systems [04]. Another example is Google's project to add 1,000 languages to Google Translate for low-resource languages [05]. Yet another is the M2M-100 project by Meta, the first multilingual MT model that translates without relying on English [06].

In summary, artificial corpora are highly beneficial in contributing to translation quality in NMT systems.

2. The PASC Project

2.1. Supervised generation

CJKI has launched a project to develop multilingual synthetic corpora, referred to as **Parallel Annotated Synthetic Corpora** or PASC, based on our comprehensive lexical databases for CJK languages and Arabic, including Arabic dialects. The goal is to provide the AI and NLP communities with a totally new large-scale series of corpora for such applications as machine learning, NMT, ASR, TTS, and deep learning, especially in domains for which rich resources are not available, such as personal names, place names, points of interest, and technical terms.

Unlike augmented corpora, which are created by enhancing or expanding existing natural corpora, we used supervised generation to create synthetic corpora from scratch using carefully crafted sentence templates based on strict conformance to rules. By following specific rules, we use templates to generate synthetic data that precisely represents real-world language patterns and structures. This means that we have full control of every phase and every feature of the corpus creation process, ensuring accurate translations, perfect bilingual or multilingual alignment, accurate grammatical and syntactic annotation, accurate phonemic and phonetic transcriptions, accurate tokenization, and more.

2.2. Initial PASC corpora

Our institute is currently engaged in developing very large-scale synthetic corpora for many domains, some of which are listed below. For example, our Japanese Personal Names Corpus (PASC-JEN) has 152 million tokens covering 1.35 million personal names, translated into approx. 18 million Chinese, Japanese and Arabic equivalents.

1. Multilingual Synthetic Corpus of Japanese Personal Names (PASC-JEN)
2. Multilingual Synthetic Corpus of Japanese Place Names (PASC-JPN)
3. Multilingual Synthetic Corpus of Worldwide Place Names (PASC-WPN)
4. Multilingual Synthetic Corpus of Arab Personal Names (PASC-DAN)

3. Distinctive Characteristics

The PASC corpora have special features not present in natural and augmented corpora:

- **Fully aligned.** Unlike other parallel corpora, PASC corpora are 100% fully bilingually (or multilingually) aligned on every level: sentential, phrasal, lexical, grammatical, syntactic and punctuational (to the extent linguistically possible)
- **Translation accuracy.** The target languages are 100% faithful, accurate and idiomatic translations.
- **Accurate Transcription.** For non-Latin scripts such as Japanese, Chinese, and Arabic, mostly 100% accurate transcriptions (e.g., hiragana, pinyin, romanization) are provided. IPA and SAMPA are also available.
- **Multilingual corpora.** PASC corpora are available in monolingual, bilingual and multilingual formats in several target languages, including major and minor languages such as English, Romance, German, Chinese, Japanese, Arabic, Vietnamese, Korean, Hebrew and others.
- **Fully annotated.** Both the source and target languages can be provided with partial or full annotation, including POS tags, syntactic tags, phonological tags, and more in any desired format. The annotation is not shown in the samples in the appendix. A rich set of annotation tags are provided customized to specific applications.
- **Consistent format.** Unlike natural corpora, synthetic corpora are constructed by following precise formatting rules and ensuring 100% consistency. Typos, electronic garbage, inconsistent formatting, corrupt encoding, and other anomalies can never occur. As a result, we possess the capability for unrestricted customization in annotation schemes and data formatting. To demonstrate this capability, we have included data in the CoNLL-U format in the appendix.

4. Summary

The PASC project offers a promising solution to the challenges faced in NLP applications for processing low-resourced domains like proper nouns. By providing fully aligned and accurate synthetic corpora with precise annotations in multiple languages, PASC aims to enhance the quality of language models, including Neural Machine Translation (NMT),

Automatic Speech Recognition (ASR), and Text-to-Speech (TTS). With the potential to substantially enhance the quality of NLP algorithms, PASC opens new avenues for research and development in multilingual artificial intelligence and deep learning domains.

5. References

1. Zhang, J., & Matsumoto, T. (2019). Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora. *Applied Sciences, 9(10),* 2019.

2. Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *arXiv*.

3. Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71-90.

4. Hassan, H., Elaraby, M., & Tawfik, A. (2017). Synthetic Data for Neural Machine Translation of Spoken-Dialects. *arXiv*.

5. Bapna, A., & Firat, O. (2019, October 11). Exploring Massively Multilingual, Massive Neural Machine Translation. *Google Research Blog*.

6. Fan, A. (2020, October 19). Introducing the First AI Model That Translates 100 Languages Without Relying on English. *Facebook Newsroom.*

APPENDIX – DATA SAMPLES

Note that the PASC corpora can be provided in any format, including multilingually aligned lexical databases. Upon request, rich annotation in the form of morphological, syntactic, and phonological tags can be provided and customized to specific applications.

Larger, multilingual, pre-tokenized samples with phonological annotation can be downloaded from our website (Excel file).

Western Names

ID	ENGLISH	JAPANESE
0002-01	My full name is [Michael Owen].	私の姓名は[オーウェン・マイケル]です。
0002-02	[Michael] is my given name and [Owen] is my surname.	[マイケル]は私の名前で、[オーウェン]は私の苗字です。
0002-03	I'm called [Michael Owen].	[オーウェン・マイケル]と言います。
0002-04	Both [Michael] and [Owen] are personal names.	[オーウェン]と[マイケル]は両方とも人名です。
0002-05	[Michael Owen] is my full name.	[オーウェン・マイケル]とは私のフルネームです。
0002-06	[Michael Owen] is what's written on my ID.	旅券に記載されている姓名は[オーウェン・マイケル]です。
0002-07	I've never heard of anyone called [Michael Owen].	[オーウェン・マイケル]と言う人のことを聞いたことがない。
0002-08	I go by the name [Michael Owen].	[オーウェン・マイケル]と言う名前で呼ばれています。
0002-09	Do you know of anyone who goes by the name of [Michael Owen]?	[オーウェン・マイケル]という人を知っていますか。

Japanese Names

ID	JAPANESE	ENGLISH
0030-01	私の姓名は[森隆大]です。	My full name is [Takahiro Mori].
0030-02	[隆大]は私の名前で、[森]は私の苗字です。	[Takahiro] is my given name and [Mori] is my surname.
0030-03	[森隆大]と言います。	I'm called [Takahiro Mori].
0030-04	[森]と[隆大]は両方とも人名です。	Both [Takahiro] and [Mori] are personal names.
0030-05	[森隆大]とは私のフルネームです。	[Takahiro Mori] is my full name.
0030-06	旅券に記載されている姓名は[森隆大]です。	[Takahiro Mori] is what's written on my ID.
0030-07	[森隆大]と言う人のことを聞いたことがない。	I've never heard of anyone called [Takahiro Mori].
0030-08	[森隆大]と言う名前で呼ばれています。	I go by the name [Takahiro Mori].
0030-09	[森隆大]という人を知っていますか。	Do you know of anyone who goes by the name of [Takahiro Mori]?

Chinese Names

ID	CHINESE	ENGLISH
0040-01	我的姓名是[张小东]。	My full name is [Xiaodong Zhang].
0040-02	[小东]是我的名字，[张]是我的姓。	[Xiaodong] is my given name and [Zhang] is my surname.
0040-03	我叫[张小东]。	I'm called [Xiaodong Zhang].
0040-04	[小东]和[张]都是人名。	Both [Xiaodong] and [Zhang] are personal names.
0040-05	[张小东]是我的姓名。	[Xiaodong Zhang] is my full name.
0040-06	我的身份证上的姓名是[张小东]。	[Xiaodong Zhang] is what's written on my ID.
0040-07	我从未听过叫[张小东]的人。	I've never heard of anyone called [Xiaodong Zhang].
0040-08	我叫[张小东]。	I go by the name [Xiaodong Zhang].
0040-09	你知道叫[张小东]的人吗？	Do you know of anyone who goes by the name of [Xiaodong Zhang]?

Korean Names

ID	KOREAN	ENGLISH
0050-01	저의 성명은 [김지영]입니다.	My full name is [Jiyeong Gim].
0050-02	[지영]은 저의 이름이고, [김]은 저의 성입니다.	[Jiyeong] is my given name and [Gim] is my surname.
0050-03	저는 [김지영]이라고 합니다.	I'm called [Jiyeong Gim].
0050-04	[지영]과 [김]은 모두 다 인명입니다.	Both [Jiyeong] and [Gim] are personal names.
0050-05	[김지영]은 저의 성명입니다.	[Jiyeong Gim] is my full name.
0050-06	저의 신분증의 이름은 [김지영]입니다.	[Jiyeong Gim] is what's written on my ID.
0050-07	[김지영]이라는 이름은 들어본 적이 없습니다.	I've never heard of anyone called [Jiyeong Gim].
0050-08	저는 [김지영]이라고 합니다.	I go by the name [Jiyeong Gim].
0050-09	[김지영]이라는 분을 아시나요?	Do you know of anyone who goes by the name of [Jiyeong Gim]?

Arabic Names

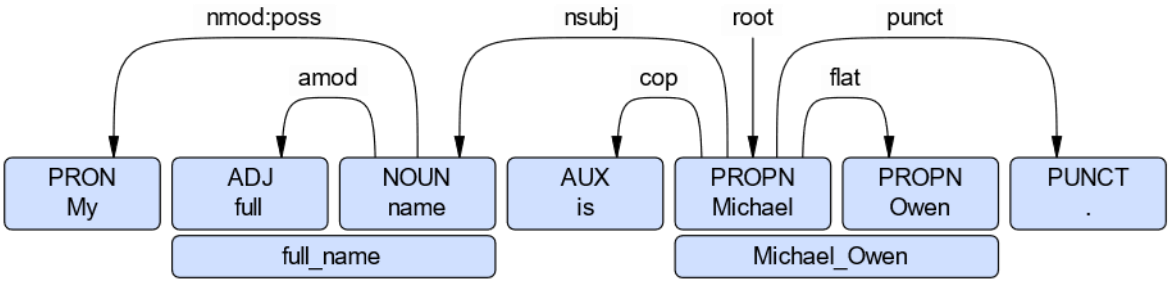
ID	ARABIC	ENGLISH
0060-01	اسمي الكامل هو [محمد العبدي].	My full name is [Mohammed Al-Abadi].
0060-02	[محمد] هو اسمي الاول، و [العبدي] هو اسمي العائلي.	[Mohammed] is my first name and [Al-Abadi] is my family name.
0060-03	أنا أدعى [محمد العبدي].	I'm called [Mohammed Al-Abadi].
0060-04	[محمد] و[العبدي] كلاهما أسماء شخصية.	Both [Mohammed] and [Al-Abadi] are personal names.
0060-05	[محمد العبدي] هو اسمي الكامل.	[Mohammed Al-Abadi] is my full name.
0060-06	الإسم المدرج في بطاقة هويتي هو [محمد العبدي].	The name listed on my ID card is [Mohammed Al-Abadi].
0060-07	لم أسمع عن أحد يدعى [محمد العبدي].	I haven't heard of anyone called [Mohammed Al-Abadi].
0060-08	أنا ألقب ب [محمد العبدي].	I go by the name [Mohammed Al-Abadi].
0060-09	هل تعرف شخصا يلقب بـ[محمد العبدي]؟	Do you know of anyone who goes by the name [Mohammed Al-Abadi]?

APPENDIX – ANNOTATION

CoNLL-U: Annotation

ID	FORM	UPOSTAG	MISC
# sent_id = en-ja-0001-01 # text = My full name is Michael Owen. # text_ja = 私の姓名はオーエン・マイクルです。			
1	My	PRON	Gloss=私の
2-3	full_name	_	Gloss=姓名
2	full	ADJ	_
3	name	NOUN	_
4	is	AUX	Gloss=は
5-6	Michael_Owen	_	Gloss=オーエン・マイクル|NamedEntity=B-PER
5	Michael	PROPN	Gloss=マイクル|NamedEntity=B-PER:FN
6	Owen	PROPN	SpaceAfter=No|Gloss=オーエン|NamedEntity=I-PER:LN
7	.	PUNCT	SpaceAfter=No

CoNLL-U: Semantic Tree Representation



ABOUT

The CJK Dictionary Institute

The CJK Dictionary Institute (CJKI) specializes in CJK and Arabic computational lexicography. The institute creates and maintains CJK (Chinese, Japanese and Korean) and Arabic lexical databases currently covering approximately 50 million entries. Located in Saitama, Japan, CJKI is headed by Jack Halpern, editor-in-chief of the world-renowned New Japanese-English Character Dictionary and of various other CJK dictionaries.

CJKI plays a leading role in helping the IT industry penetrate the lucrative East Asian market by providing software developers with high quality dictionary data. This includes comprehensive databases of general vocabulary, proper nouns, and technical terms for CJK languages, including Chinese dialects such as Cantonese and Hakka. CJKI also maintains databases and romanization systems of Arabic proper nouns, a large-scale Spanish-English dictionary, and various multilingual databases of proper nouns and geographic data.

CJKI has become one of the world's prime sources for CJK lexical resources. It is contributing to CJK and Arabic information processing technology by providing high-quality lexical resources and professional consulting services to some of the world's leading software developers and IT companies, including Fujitsu, Sharp, Sony, IBM, Google, Microsoft, Yahoo, Amazon, and Baidu.

Jack Halpern

Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, is a lexicographer by profession. For sixteen years was engaged in the compilation of the New Japanese-English Character Dictionary, and as a research fellow at Showa Women's University (Tokyo), he was editor-in-chief of several kanji dictionaries for learners, which have become standard reference works.

Jack Halpern, who has lived in Japan over 40 years, was born in Germany and has lived in six countries including France, Brazil, Japan, and the United States. An avid polyglot who specializes in Japanese and Chinese lexicography, he has studied 18 languages (speaks 11 fluently) and has devoted several decades to the study of linguistics and lexicography.

On a lighter note, Jack Halpern loves the sport of unicycling. Founder and long-time president of the International Unicycling Federation, he has promoted the sport worldwide and is a director of the Japan Unicycling Association. Currently, his passions are playing the quena and improving his Chinese, Esperanto, and Arabic.