JSAI Technical Report
SIG-AGI-014-010

Bacteria Lingualis on BERToids – – Concept Expansion for Cognitive Architectures

Rafal Rzepka[1,2*] Sho Takishita[1,2] Kenji Araki[1]

[1] Faculty of Information Science and Technology, Hokkaido University
[2] RIKEN Center for Advanced Intelligence Project (AIP)

Abstract: In this paper, we introduce our trials to extend our previous work on life-long learning cognitive architecture Bacterium Lingualis, which collects world knowledge from textual resources using linguistic clues. We utilize mask prediction functionality of the BERT language model to augment simple concepts with additional knowledge like means, goals or merits and demerits. We present results of preliminary tests of the additional knowledge in an automatic ethical assessment task and report our findings.

1 Introduction

Creating large simulation environments for cognitive architectures is costly and time-consuming, therefore such architectures are tested in simple mazes, with limited objects, robotic functions, or objectives. However, when it comes to the real world, sudden and unpredictable changes of an environment or agents within it, are practically unavoidable. To deal with a rich context, an agent needs to not only possess detailed knowledge but also be able to calculate how a newly received signal can cause problems for executing a current plan or change user's goals or preferences.

In our previous work, we have shown how augmentation can be performed in the task of creating longer event chains [Takishita 19]. During the experiments, we confirmed that controlling the flow of additional knowledge is not an easy task, and for example, it is not obvious if a place of an act changes in every step or remains the same. We concluded that to tackle such contextual changes, an agent must posses additional, often tacit, knowledge, and an ability to estimate how the semantic elements of its influence the future, e.g., user's well-being. In this paper, we a) propose a method for simple automatic addition of context information and b) investigate how the richness of context influences human's and machine's ability to asses emotions conveyed by an act.

*Contact:
Language Media Laboratory
Faculty of Information Science and Technology
Hokkaido University
Kita-ku, Kita 9, Nishi 14, Sapporo, Japan
E-mail: rzepka@ist.hokudai.ac.jp

2 Augmenting Common Sense

As we described in our previous work [Rzepka 19], natural language processing can be useful for advancing artificial intelligence. Still, the lack of common sense knowledge remains unsolved due to, among other factors, the problem of tacit knowledge, pieces of information which are so apparent for human beings, that are left unspoken or unwritten. Because what is not stated tends to be omitted by machine learning algorithms, in our research, we try to artificially add missing knowledge to enrich the machine's ability to simulate possible scenarios. Our basic approach was to model a simplistic micro cognitive architectures [Rzepka 03] "living" on the Web and, for instance, learn semantic categorization using Japanese particles in an unsupervised manner. In this trial, we extend this knowledge completion by replacing the previous simple corpus search approach with more context-dependent language model BERT [Devlin 19].

2.1 Related Work

Automatic common sense knowledge acquisition from big data is studied since the dawn of the Internet era. Projects such as YAGO [Suchanek 07], NELL [Carlson 10] or WebChild [Tandon 14] aim at extracting semantic assertions from unstructured text data found on the Internet, however most of the other harvesting

methods [Weikum 19] aim at the discovery of existing relations, not at artificially augmenting the discovered knowledge with what was left unsaid (unwritten). A commonsense knowledge harvesting method using a language model has been recently presented by [Rajani 19]. Their Commonsense Auto-Generated Explanation (CAGE) framework uses the large OpenAI GPT [Radford 18] language model, which is fine-tuned on common sense question-answering datasets to explain-and-then-predict (reasoning) and predict-and-then-explain (rationalization) tasks. Although the accuracy improvement was high (6-10 points depending on a task), its implementation is impossible for languages other than English. Unlike our agent, which is fully unsupervised, CAGE requires manual data preparation and does not tackle with contexts changes. On the other hand, recently created "Mosaic" knowledge graphs ATOMIC [Sap 19], and COMeT [Bosselut 19] (also only for English) are probably the closest knowledge bases to be able to deal with contextual changes, but they are costly to create.

There are also trials with adding embeddings-based knowledge from a text in real-world applications similar to increase machine's commonsense reasoning capabilities. One of the most recent ones is RoboCSE (Robot Common Sense Embedding) [Daruna 19], which showcases how multi-relational embeddings can be leveraged in robotics to model semantic knowledge about object affordances, locations, and materials slightly resembling our previous work with robots [Takagi 11]. Although representing promising direction, their system is limited to a few items. To the authors' best knowledge, implementing neural language modeling in cognitive architectures remains hypothetical [Santhanam 19].

2.2 Targeted Knowledge Types

In this research, we aim at testing semantic categories that are less popular or non-existent in classic knowledge acquisition approaches. For the preliminary tests, we chose eight topics we thought might be useful for cognitive architectures to gather information about to store it in their long-term memory for future analysis of changing contexts. *Goal* is meant for guessing reasons of acts A_1 and A_2. *Place* is for predicting its common locations. *Mean* represents means or tools which are used in A_1. *Companion* is for retrieving people and objects that accompany an act. *Difficulty* is meant for guessing what kind of human incapabilities can be prerequisites for A_1 to be followed by A_2. *Cost* is for checking if BERT is able to predict monetary value related to a given pair of acts. *Problem* is supposed to consist of words indicating possible problems following the acts. And *Merit* type is to gather benefits when A_1 and A_1 are combined.

3 Augmentation Algorithm

The knowledge augmentation process is shown in Figure 1. As an input, we use pairs of simple events in causal relation retrieved from approximately 100 million Japanese Web pages [Shibata 14]. They are generated by calculating the co-occurrence measure between predicate-arguments, applying association rule mining, and identifying arguments by implementing case frames. The database contains approximately 100,000 unique events, with approximately 340,000 strongly-related event pairs. An input pair from the database is divided into left and right acts A_1 and A_2 stripped to core verbs in order to investigate the language model's ability to add further words. To achieve it we used the standard version of Japanese BERT [Shibata 19] and by using Japanese articles *ga* for a grammatical subject, *wo* for object and *de* for places and tools are we gathered preceding words (25 most probable candidates) with the masking phrase [MASK]+particle+A. Acquired words were then filtered by JUMAN++ morphological analyzer[1] and if the word was a noun with an existing JUMAN's semantic category it was then searched in YACIS corpus [Ptaszynski 12] for occurrences. If the extended act did not exist in the corpus, it was discarded. We have noticed that BERT's accuracy drops after an unknown (not in vocabulary) word is predicted. Therefore the system ignored all word candidates after detecting unknown output.

The candidates were then used in sentiment-category related phrases to create masked sentences for BERT to achieve additional knowledge in categories described in the previous section:

- Goal: [MASK] のために+A_1+ことによって+A_2+ことができる (in order to [MASK] + (one can) + A_1 + and thanks to it + A_2+becomes possible)
- Place: [MASK] という場所で + A_1 +ことが起きて， + A_2+のは当たり前 (at the place of

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

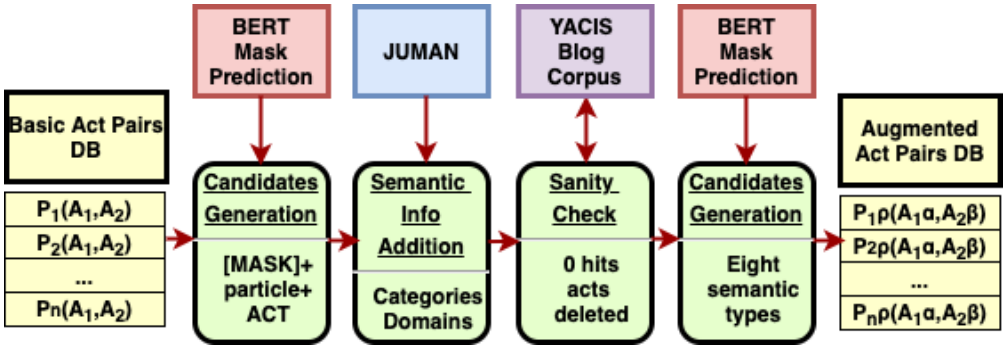

Figure 1: Knowledge augmentation process.

[MASK]+A_1+happened and+A_2+was obvious)

- Mean: [MASK] を使って+A_1+後で+A_2+ことになる (A_1+after using [MASK]+it will+A_2)
- Companion: [MASK] と一緒に+A_1+ ことになったら +A_2+ でしょう (if A_1 (happens) together with [MASK]+it will+A_2+won't it?)
- Difficulty: [MASK] できなくて+A_1+ ことになったら +A_2+ ことになるでしょう (if (one) can't do [MASK] and+A_1+happens+A_2+ will surely happen)
- Cost: A_1+としたら+A_2+のは [MASK] 円かかる (if one A_1+and+A_2+it will cost [MASK] yen)
- Problem: A_1+後で+A_2+のは [MASK] という問題につながってしまう (A_1+after で+A_2+will lead to the problem of [MASK])
- Merit: A_1+後で+A_2+のは [MASK] というメリットがあります (if one A_1+and then+A_2+it will lead to a merit of [MASK])

The phrases were created heuristically to fit the dictionary forms of the acts and represent various modalities for further investigation of their effects.

In the following section, we show examples of the words automatically assigned to proposed types and describe our experiment for the context influence of affective change in augmented knowledge.

4 Experiments and Results

4.1 Category Agreement

As our knowledge augmentation process is on-going, we have examined only a part of the event database developed by [Shibata 14]. We have taken 5,769 verb pairs which have been augmented so far and the first author visually checked the agreement between words in an estimated category and the JUMAN's category (see Table 1. Except few problematic entries like suffix 機 (for machinery), word 無理 (impossibility) and adjective よい (good) which usually belong to the preceding words, we considered the data sufficient for further experimentation.

4.2 Context Influence on Sentiment

4.2.1 Experimental Data Creation

Evaluating any kind of common sense knowledge is difficult, as everyone's background and experiences are challenging. Our assumption is that the more a person knows about a context, it is easier to assess the commonness of a given situation. We also hypothesize that such knowledge can be evaluated more precisely when used in a task, not just when shown to human subjects. As we are interested in Artificial Moral Agents, we design a simple experiment to see how the context influences agent's performance in estimating an emotional load of an act and how the human evaluators interpret shorter and longer acts. To create a set of relatively similar acts, we chose Japanese verb 殴る ("to punch" or "to hit") and let our system semi-randomly generate sentences with newly acquired words. For increasing the naturalness of sentences, we added agents and patients (covered in other ongoing research) retrieved by phrases [MASK]+という人が+A and [MASK]+という人を+A (indicating people: doers and receivers of the act A). After eliminating similar phrases, manually exchanging places and patients for increased variety, we have chosen 50 sentences with various lengths, places, means

Table 1: Translation of the top augmenting words and the corresponding top categories and domains of JUMAN

Types	Words *(top 5)*	Category *(top 1)*	Domain *(top 1)*
Goals	goal, living, person, child, family	Abstract	Family & Life
Places	town, village, station, hill, park	Place (facility)	Administration
Means	machine (suffix), time, phone, car, ship	Artifacts (other)	Transport
Companions	person, I, family, child, friend	Person	Family & Life
Difficulties	participation, satisfaction, travel, lodging, marriage	Abstract	Recreation
Costs	100, 10000, 10, 1, 2, 5, 500, 1000	Abstract	None
Problems	time, impossibility, country, marriage, family	Abstract	Family & Life
Merits	time, free, impossibility, comfort (easiness), good	Abstract	Business

(tools) and objects. For example: 壁を殴る ("punches a wall"), 道路という場所で人を殴る ("hits a man on a street") or 力士が手を使って土俵という場所で子供を殴る ("sumo wrestler hits a child with his hand on the sumo wrestling ring").

4.2.2 Sentiment Analysis

As the lexicon-based approaches are not meant for longer sentences, we utilized stochastic approaches. First, with a plan to leave all the processes to BERT without fine-tuning, we have used its next-sentence prediction (abbreviated to NSP from now on) functionality and introduced a simplistic method of adding a positive and negative phrases as a follow-up of the input. The positive one was input+ことになって+最高の結果となりました (input+ "happened and it lead to the best ever result") and the negative one was identical apart of word 最高 (the best, greatest) was replaced by 最悪 (the worst, lousiest). Again, the phrases were chosen intuitively without much experimenting. To compare this approach with a more conventional method, we also used Asari, a sentiment analyzer for Japanese language based on tf-idf vectorizer and linear SVC[2]. The sentiment of NSP was calculated from a difference between probabilities of both phrases; for Asari we have utilized its output score as it is (it was negative for all inputs).

4.2.3 Findings from Results

We have asked 8 university students (7 males and 1 female) to estimate how unethical are the 50 generated sentences on a scale from 0 to 10 after reading them first to see the differences. Then we normalized the ranges outputs of NSP and Asari to correspond to the 0-10 scale.

When compared to human evaluators, NSP and Asari differed in their performance in the way that the latter had more perfect matches (9/50, whereas the former has agreed only once, see Figure 2), but both had similar overall standard error rate (NSP: 0.43 vs. Asari: 0.40).

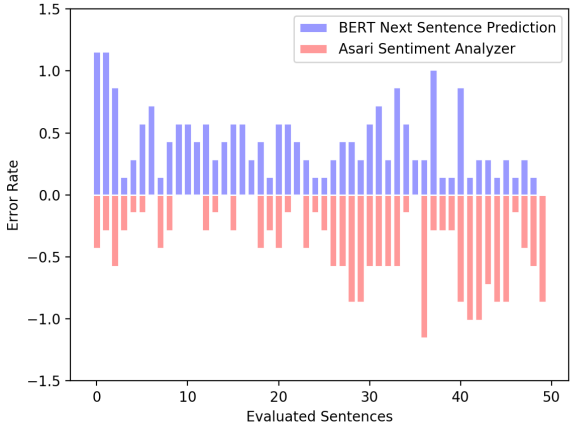

Figure 2: Error rate comparison between utilized sentiment predictions (Kolmogorov-Smirnov test p-value = 0.154). Asari results inverted for better visibility.

Because there was a very slight agreement between evaluators (see Table 2, we calculated the accuracy in three levels of strictness (see Table 3, but even when the difference between human average evaluation was set to treat two neighboring points as correct, both methods barely exceeded 50.0% of correctness. As can be seen in Table 2, we also checked how the agreement between evaluators has altered. Although the data is too small to draw convincing conclusions,

[2] https://github.com/Hironsan/asari

Table 2: Agreement between evaluators depending on the context richness.

Metric	All	Longest	Shortest	Mid-Val
Kappa	0.072	0.068	0.048	0.099
Fleiss κ	0.071	0.067	0.044	0.096
Alpha	0.064	0.060	0.031	0.091
Scott's π	0.061	0.052	0.023	0.083

Table 3: Accuracy rates depending on the agreement strictness.

	BERT's NSP	Asari Analyzer
Strict	0.02	0.18
$\pm$1 error	0.24	0.32
$\pm$2 error	0.50	0.52

the shortest 15 sentences indicated a small agreement, and the longest 15 showed a greater degree. However, the sentences of the middle length appeared to have the highest agreement.

When it comes to the automatic sentiment evaluation, the error rate dropped for both systems 0.3 points between the shortest and longest sentences, suggesting that more knowledge can be helpful in estimating the negativeness of the experimental dataset. We also allowed evaluators to mark sentences that were difficult to interpret. Two sentences were assessed as hard to comprehend by half or more subjects: ゲームのためにホールという場所で魔法を使ってプレイヤーがキャラクターを殴る ("player hits a character at the hall with a spell") and ボタンを使ってキャラクターを殴る ("hit a character with a button"). It seems that generated sentences related to video games caused some confusion, and this problem indicates that some other set, preferably limited to more concrete situations, would be a better choice for future tests.

5 Conclusions and Future Work

In this paper, we have described how the knowledge of a cognitive architecture can be enriched by a neural language model which considers context in a deeper manner than its predecessors. Our preliminary experiments showed that it is possible to augment simple knowledge with missing elements by merely creating a single phrase capturing words specific to a given category. This approach also shows sentiment prediction functionality similar to a non-neural tool. However, there are many remaining problems to make this approach useful, e.g., for the moral decision making, and the scale and setup of our preliminary experiments are not yet sufficient to draw strong conclusions.

Obviously there is a need for extending the current simple current acts into richer ones and mechanizing the process of generating queries for existing and new semantic categories. Better phrases for both augmenting and sentiment prediction methods can be found, hence it is necessary to investigate automatic approaches, e.g., by combining neural approach and lexicon-based methods, distant supervision or bootstrapping. Deep learning is known for its limited reasoning, but we believe it is able to provide relatively rich string of contextual signals to simulate real world situations without costly data preparation.

Most recent achievements of language models, as REALM by Google researchers (Retrieval-Augmented Language Model Pre-Training"[3]) suggest that the trend of mixing them with retrieved, more organized or controlled data can be the next step toward more useful, possibly unsupervised, life-long learners like Bacteria Lingualis and other, more sophisticated cognitive architectures.

6 Acknowledgments

This work was supported by JSPS Kakenhi Grant Number 17K00295.

References

[Bosselut 19] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y.: COMeT: Commonsense transformers for automatic knowledge graph construction, *arXiv preprint arXiv:1906.05317* (2019)

[Carlson 10] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E., and Mitchell, T.: Toward an Architecture for Never-Ending Language Learning, in *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)* (2010)

[Daruna 19] Daruna, A., Liu, W., Kira, Z., and Chetnova, S.: RoboCSE: Robot common sense

[3] https://realm.page.link/paper

embedding, in *IEEE International Conference on Robotics and Automation (ICRA 2019)*, pp. 9777–9783 (2019)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Ptaszynski 12] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., and Momouchi, Y.: YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information, in *Proceedings of The AISB/IACAP World Congress*, pp. 40–49 (2012)

[Radford 18] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.: Improving language understanding by generative pre-training, `https://openai.com/blog/language-unsupervised` (2018)

[Rajani 19] Rajani, N. F., McCann, B., Xiong, C., and Socher, R.: Explain Yourself! Leveraging Language Models for Commonsense Reasoning, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy (2019), Association for Computational Linguistics

[Rzepka 03] Rzepka, R., Araki, K., and Tochinai, K.: Bacterium Lingualis – The Web-Based Commonsensical Knowledge Discovery Method., in Grieser, G., Tanaka, Y., and Yamamoto, A. eds., *Discovery Science*, Vol. 2843 of *Lecture Notes in Computer Science*, pp. 460–467, Springer (2003)

[Rzepka 19] Rzepka, R., Takishita, S., and Araki, K.: Unicorn Story Generation and Limits of Words – On Perspectives of Automatic Tacit Knowledge Addition, Technical report, Technical Report of JSAI Special Interest Group for Artificial General Intelligence, SIG-AGI-011-10 (2019)

[Santhanam 19] Santhanam, S. and Shaikh, S.: A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions, *ArXiv*, Vol. abs/1906.00500, (2019)

[Sap 19] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y.: ATOMIC: An atlas of machine commonsense for if-then reasoning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3027–3035 (2019)

[Shibata 14] Shibata, T., Kohama, S., and Kurohashi, S.: A Large Scale Database of Strongly-related Events in Japanese, in *LREC* (2014)

[Shibata 19] Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S.: Improving Parsing Accuracy of Japanese Using BERT (in Japanese), in *Proceedings of The 25th Anual Conference of Association for Linguistic Processing (ANLP), pp.205-208* (2019)

[Suchanek 07] Suchanek, F. M., Kasneci, G., and Weikum, G.: Yago: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, in *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 697–706, New York, NY, USA (2007), ACM

[Takagi 11] Takagi, K., Rzepka, R., and Araki, K.: Just Keep Tweeting, Dear: Web-Mining Methods for Helping a Social Robot Understand User Needs, in *Proceedings of Help Me Help You: Bridging the Gaps in Human-Agent Collaboration*, pp. 60–65, Symposium of AAAI 2011 Spring Symposia (SS-11-05) (2011)

[Takishita 19] Takishita, S., Rzepka, R., and Araki, K.: Implicit Knowledge Completion Method Using Relevance Calculation of Distributed Word Representations, in *Bridging the Gap Between Human and Automated Reasoning Workshop at IJCAI 2019, Macao* (2019)

[Tandon 14] Tandon, N., De Melo, G., Suchanek, F., and Weikum, G.: Webchild: Harvesting and organizing commonsense knowledge from the web, in *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 523–532 (2014)

[Weikum 19] Weikum, G., Hoffart, J., and Suchanek, F.: *Knowledge Harvesting: Achievements and Challenges*, pp. 217–235, Springer International Publishing, Cham (2019)