Conversational text composition through commonsense knowledge

Angel Rendon[0000−0003−3900−9582]

Universidad Nacional de Colombia
amrendonsa@unal.edu.co
http://unal.edu.co/

Abstract. Natural Language Processing techniques allows us to process text in wide range ways, making possible to extract key information out of texts and even proposing machine translators systems. One of those possibilities is tied to having well trained systems to have smart enough conversations with humans. This work aims to analyze the state-of-the-art techniques and implement them in the construction of a system that using different methods, make it possible to sustain a basic conversation on general topics.

Keywords: commonsense knowledge · natural language processing · machine learning · semantic association

1 Introduction

One of the artificial intelligence (AI) keystones would be definitively be having fully conversational systems to interact with people for several applications ranging from recommender systems, expert systems, to specialized chatbots and assistants [1].

Several techniques based on Machine Learning (e.g. Bayesian models, SVM, supervised and unsupervised learning methods), and statistical model methods (e.g. word frequency, text rank, and inverse document frequency), have been used for a long time, with promising results.

However, systems based on these techniques rely on well formed corpora. As an example, WordNet [5] has the following synset for *cat*:

```
S: (n) computerized tomography, computed tomography, CT,
computerized axial tomography, computed axial tomography,
CAT (a method of examining body organs by scanning them
with X rays and using a computer to construct a series of
cross-sectional scans along a single axis)
```

This simple example could lead to think that it would be possible to miss that specific synset when talking about *computerized tomography* when using

[1] https://www.youtube.com/watch?v=d40jgFZ5hXk

cat in a medical text, showing instead the most probable definition *feline*. That scenario is plausible since the knowledge is strongly dependent on the quality of either unstructured texts or its scale and domain-specific knowledge [4].

Lexical semantic understanding, sustained by socially shared commonsense knowledge on the core as it has the content of what people intends to know while conversing [7].

This contribution aims to:

1. integrates existing work on the semantic association for construction of commonsense knowledge.
2. build an English conversational system using semantic association based on commonsense knowledge construction techniques.
3. assess the performance of the built system compared to traditional conversational approaches.

2 Related work

A key element to enrich understanding of sentences is tied to word senses. This fact is more important when the word belongs to a specific domain context. Through hybrid clustering techniques [6] chooses the best terms elements to build the initial committees (i.e. clusters of meanings for a word) out of the WordNet corpus [5], and removing the sense so through further iterations the algorithm can effectively analyze other synsets.

State-of-the-art pre-trained models capture commonsense knowledge with limited value for domain-specific contexts. [3] proposes a general model for capturing specific domain knowledge of software engineering through a word2vec model exploting available information on Stack Overflow posts. Their work shows that the model effectively desambiguates metaphorical use of English words when it comes to this specific domain, capturing well grained relations within it.

Distributional Semantic Models are another technique for supporting semantic understanding for natural language processing. These models are highly dependent on the size and quality of the corpora that has the commonsense knowledge for the comprehensive task. While English do have high quality and large scaled commonsense and domain specific information, other languages lack enough material to build comprehensive distributional models. [1] proposes to combine lightweight machine translation model using the English Distributional Semantic Model for building enriched knowledge word vectors for other languages. By building a model leveraged by a unigram-level source-target probabilities which can be directly computed from the parallel corpora, the author gets word vectors for other languages from English DSM as feasible activity, showing up that about 66% of improvement has been gotten from lightweight models compared to other models. Spanish got the best performance with nearly 60%, whereas Dutch got the worst with nearly 50%. This demonstrates that lightweight machine translation models is, in the worst case, equivalent (in some cases outperforming) to the state-of-the-art machine translation systems for translation of word pairs.

As stated at the beginning, one of the important milestones on AI is having intelligent systems to perform common activities as an agent spawned at any location performing a set of actions to answer a question. These actions might require the agent to navigate, process images, understand natural language and learn. [2] proposes a new AI task, where an agent is dropped anywhere and must navigate processing images to gather as much information to answer questions about the environment.

References

1. Barzegar, S., Davis, B., Handschuh, S., Freitas, A.: Multilingual Semantic Relatedness Using Lightweight Machine Translation. Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018 **2018-Janua**, 108–114 (2018). https://doi.org/10.1109/ICSC.2018.00024
2. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied Question Answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10. IEEE (jun 2018). https://doi.org/10.1109/CVPR.2018.00008, https://ieeexplore.ieee.org/document/8578106/
3. Efstathiou, V., Chatzilenas, C., Spinellis, D.: Word embeddings for the software engineering domain. In: Proceedings of the 15th International Conference on Mining Software Repositories - MSR '18. pp. 38–41. ACM Press, New York, New York, USA (2018). https://doi.org/10.1145/3196398.3196448, http://dl.acm.org/citation.cfm?doid=3196398.3196448
4. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A Knowledge-Grounded Neural Conversation Model pp. 5110–5117 (2017), http://arxiv.org/abs/1702.01932
5. Miller, G.A., Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: WordNet: An on-line lexical database. INTERNATIONAL JOURNAL OF LEXICOGRAPHY (1990)
6. Pantel, P., Lin, D.: Discovering word senses from text **41**, 613 (2004). https://doi.org/10.1145/775047.775138
7. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X.: Commonsense knowledge aware conversation generation with graph attention. IJCAI International Joint Conference on Artificial Intelligence **2018-July**, 4623–4629 (2018)