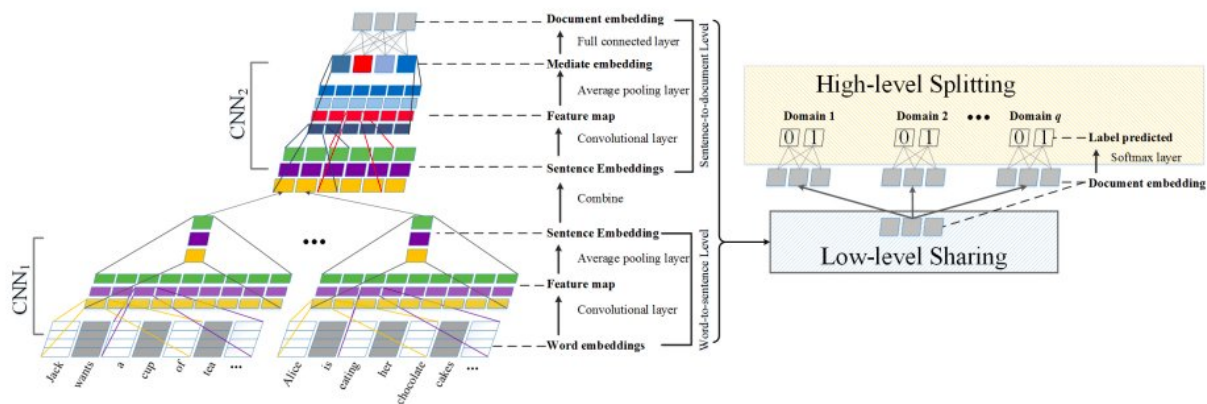


# A neural network to extract knowledgeable snippets and documents

September 5 2018, by Ingrid Fadelli



Structure of proposed model. Credit: Zhou et al.

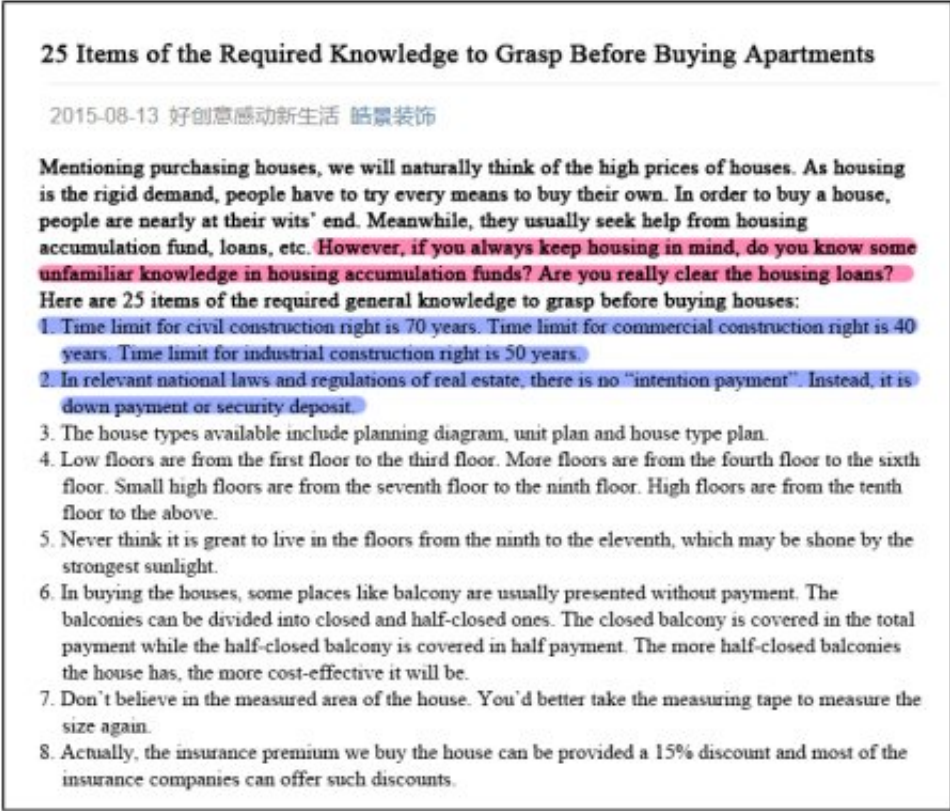
Every day, millions of articles are published on social media and other platforms, receiving a vast amounts of clicks and shares from users navigating the web. Many of these articles contain useful information that, if extracted, could be used to compile knowledge databases or to deliver knowledge retrieval and question answering services.

Researchers at the Chinese Academy of Sciences (CAS) have developed a [convolutional neural network](#) (CNN)-based model to extract knowledgeable snippets and annotate documents. Their method, outlined on a paper pre-published on arXiv, was found to perform better than

existing tools, despite being trained for shorter periods of time.

In their paper, the researchers define the term "knowledgeable document" as "a document containing multiple knowledgeable snippets, which describe concepts, properties of entities, or the relations among entities." So far, most knowledge bases, such as YAGO or DBpedia, extract knowledge based on Wikipedia, WordNet, GeoNames, and other online resources. However, compared to [social media](#) platforms, these resources often contain limited and inflexible information.

"Another recent knowledge base, Probase, with 2.7 million concepts, was automatically harnessed from the so-far largest corpus, consisting of 326 million knowledgeable sentences extracted from 1.68 billion web pages," the researchers wrote in their paper. "However, these sentences are extracted only by the Hearst patterns. For extracting more knowledgeable snippets to construct more comprehensive knowledge bases, semantic-based methods are needed to complement the previous pattern-based ones."



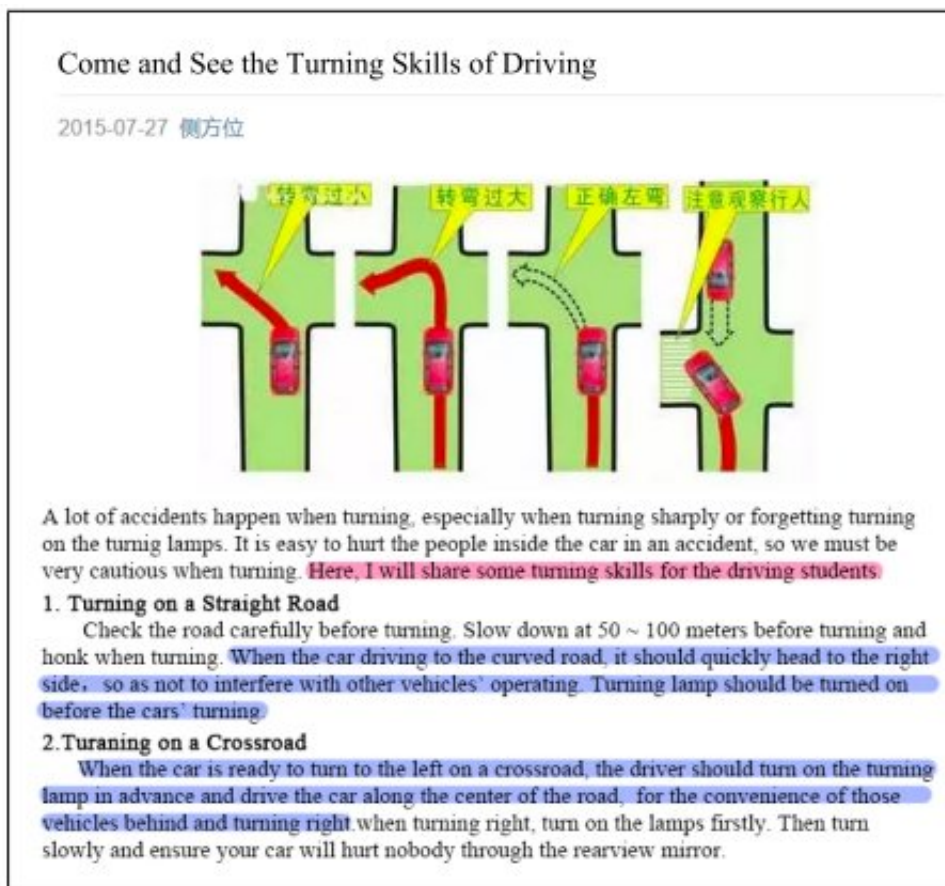
Example of knowledgeable document. The blue and red sentences are knowledgeable and unknowledgeable snippets respectively. The document introduces the 25 pieces of tips for purchasing real estates. Credit: Zhou et al.

Knowledgeable snippets and articles could also be used to develop knowledge retrieval and question answering services. These services would, for instance, answer questions raised by users who are looking for help with a particular problem. With these applications in mind, the researchers at CAS set out to develop a CNN based model that can analyze the semantics of a document, determine whether it is knowledgeable or not, and extract knowledgeable snippets of information from it.

"Specifically, we propose SSNN, a joint CNN-based model, to

understand the abstract concept of documents in different domains collaboratively and judge whether a [document](#) is knowledgeable or not," the researchers explain in their paper. "In more detail, the network structure of SSNN is 'low-level Sharing, high-level Splitting,' in which the low-level layers are shared for different domains while the high-level layers beyond the CNN are trained separately to perceive the differences of different domains."

The model devised by the researchers offers an end-to-end solution to annotate documents that does not entail extensive and time-consuming feature engineering. They also developed manual features and trained a SVM classifier model to complete the task.



Example of knowledgeable document. The blue and red sentences are

knowledgeable and unknowledgeable snippets respectively. The document introduces the turning skills of driving. Credit: Zhou et al.

The researchers evaluated the effectiveness of their model on a dataset of real documents from three content domains on WeChat, a Chinese messaging, social media and mobile payment platform developed by Tencent. Their findings were very promising, with the SSNN performing consistently better than other CNN models, while saving time and memory consumption thanks to shorter and more efficient training processes.

"Compared with building multiple domain-specific CNNs, this joint model not only critically saves training time, but also improves the prediction accuracy visibly," the researchers wrote in their paper. "The superiority of the proposed model is demonstrated in a real dataset from Wechat public platforms."

In future, the SSNN [model](#) proposed in this study could be used to build more comprehensive knowledge databases. It could also aid the development of innovative services that answer user queries both quickly and exhaustively in real-time.

**More information:** Hierarchical Neural Network for Extracting Knowledgeable Snippets and Documents. arXiv:1808.07228v1 [cs.CL]. [arxiv.org/abs/1808.07228](https://arxiv.org/abs/1808.07228)

## Abstract

In this study, we focus on extracting knowledgeable snippets and annotating knowledgeable documents from Web corpus, consisting of the documents from social media and We-media. Informally, knowledgeable snippets refer to the text describing concepts, properties

of entities, or relations among entities, while knowledgeable documents are the ones with enough knowledgeable snippets. These knowledgeable snippets and documents could be helpful in multiple applications, such as knowledge base construction and knowledge-oriented service.

Previous studies extracted the knowledgeable snippets using the pattern-based method. Here, we propose the semantic-based method for this task. Specifically, a CNN based model is developed to extract knowledgeable snippets and annotate knowledgeable documents simultaneously. Additionally, a "low-level sharing, high-level splitting" structure of CNN is designed to handle the documents from different content domains. Compared with building multiple domain-specific CNNs, this joint model not only critically saves the training time, but also improves the prediction accuracy visibly. The superiority of the proposed method is demonstrated in a real dataset from Wechat public platform.

© 2018 Tech Xplore

Citation: A neural network to extract knowledgeable snippets and documents (2018, September 5) retrieved 27 April 2024 from

<https://techxplore.com/news/2018-09-neural-network-knowledgeable-snippets-documents.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.