Title: Cache language model

URL: https://en.wikipedia.org/wiki/Cache_language_model

PageID: 31139924

Categories: Category:Computational linguistics, Category:Language modeling, Category:Machine translation, Category:Natural language processing, Category:Speech processing, Category:Speech recognition

-----

A cache language model is a type of statistical language model . These occur in the natural language processing subfield of computer science and assign probabilities to given sequences of words by means of a probability distribution . Statistical language models are key components of speech recognition systems and of many machine translation systems: they tell such systems which possible output word sequences are probable and which are improbable. The particular characteristic of a cache language model is that it contains a cache component and assigns relatively high probabilities to words or word sequences that occur elsewhere in a given text. The primary, but by no means sole, use of cache language models is in speech recognition systems. [citation needed]

To understand why it is a good idea for a statistical language model to contain a cache component one might consider someone who is dictating a letter about elephants to a speech recognition system. Standard (non-cache) N-gram language models will assign a very low probability to the word "elephant" because it is a very rare word in English . If the speech recognition system does not contain a cache component, the person dictating the letter may be annoyed: each time the word "elephant" is spoken another sequence of words with a higher probability according to the N-gram language model may be recognized (e.g., "tell a plan"). These erroneous sequences will have to be deleted manually and replaced in the text by "elephant" each time "elephant" is spoken. If the system has a cache language model, "elephant" will still probably be misrecognized the first time it is spoken and will have to be entered into the text manually; however, from this point on the system is aware that "elephant" is likely to occur again – the estimated probability of occurrence of "elephant" has been increased, making it more likely that if it is spoken it will be recognized correctly. Once "elephant" has occurred several times, the system is likely to recognize it correctly every time it is spoken until the letter has been completely dictated. This increase in the probability assigned to the occurrence of "elephant" is an example of a consequence of machine learning and more specifically of pattern recognition .

There exist variants of the cache language model in which not only single words but also multi-word sequences that have occurred previously are assigned higher probabilities (e.g., if "San Francisco" occurred near the beginning of the text subsequent instances of it would be assigned a higher probability). [citation needed]

The cache language model was first proposed in a paper published in 1990, [1] after which the IBM speech-recognition group experimented with the concept. The group found that implementation of a form of cache language model yielded a 24% drop in word-error rates once the first few hundred words of a document had been dictated. [2] A detailed survey of language modeling techniques concluded that the cache language model was one of the few new language modeling techniques that yielded improvements over the standard N-gram approach: "Our caching results show that caching is by far the most useful technique for perplexity reduction at small and medium training data sizes". [3]

The development of the cache language model has generated considerable interest among those concerned with computational linguistics in general and statistical natural language processing in particular: recently, there has been interest in applying the cache language model in the field of statistical machine translation. [4]

The success of the cache language model in improving word prediction rests on the human tendency to use words in a "bursty" fashion: when one is discussing a certain topic in a certain context, the frequency with which one uses certain words will be quite different from their frequencies when one is discussing other topics in other contexts. The traditional N-gram language models, which rely entirely on information from a very small number (four, three, or two) of words preceding the word to which a probability is to be assigned, do not adequately model this "burstiness". [citation needed]

Recently, the cache language model concept – originally conceived for the N-gram statistical language model paradigm – has been adapted for use in the neural paradigm. For instance, recent work on continuous cache language models in the recurrent neural network (RNN) setting has applied the cache concept to much larger contexts than before, yielding significant reductions in perplexity. [5] Another recent line of research involves incorporating a cache component in a feed-forward neural language model (FN-LM) to achieve rapid domain adaptation. [6]

See also

Artificial intelligence

History of natural language processing

History of machine translation

Speech recognition

Statistical machine translation

References

Further reading

Jelinek, Frederick (1997). Statistical Methods for Speech Recognition . The MIT Press . ISBN 0-262-10066-5 . Archived from the original on 2011-08-05 . Retrieved 2011-09-24 .