symptoms. Thus, apart from speech-language pathologists and clinicians, there is a major role for linguists to be played in clarifying, describing, diagnosing, assessing and providing intervention to the disorders. Thus, clinical linguistics can be introduced into the curriculum of linguistic sciences degrees.

References

1. Routledge A. D. A Glossary of Applied Linguistics / A. D. Routledge. – New York, 2016. – 152 p.

2. Schmitt N. Applied Linguistics / N. Schmitt. – New York, 2014. – 189 p.

3. Richards J. C. Longman Dictionary of Language Teaching and Applied Linguistics / J. C. Richards, R. Schmidt. – London, 2002. – 28 p.

4. Crystal D. Aspects of clinical linguistic theory and practice / Crystal. – Perth, 1981. – 25p.

5. Crystal D. Clinical linguistics. The Handbook of Linguistics / Crystal. – Oxford: Blackwell, 2001. – 673 p.

Alina Karpenko
Vasyl` Stus Donetsk National University
Vinnytsia
Research Supervisor: O.O. Odintsova, Senior Lecturer
Language supervisor: O.O. Odintsova, Senior Lecturer

PROBABILISTIC NATURAL LANGUAGE PROCESSING : BAYES THEOREM IN LANGUAGE MODELING

Introduction. Probabilistic reasoning is very important for NLP. Suppose that we have a system that recognizes speech, which converts an audio signal into text. Most of the time it will not be able to find the perfect interpretation of a speech signal. It may come up with a number of alternatives, some of which are more reasonable than others. For example, if you say “recognize speech”, it’s very possible that your system was going to hear something like “reach a crew peach”. Because for our speech recognition system those two strings sound very similar and they maybe very easy to confuse. But obviously for human being they are very different and one of them is reasonable, the other one is completely nonsensical. Which would suggest that we want the probability of the first string to be very high, and the probability of the second string to be relatively low. So, even if the speech recognition system has to chose between those two, it will have an easy time figuring out which one is correct.

Probabilistic modeling of NLP includes document clustering, topic modeling, language modeling, part-of-speech induction, parsing and grammar induction, word segmentation, word alignment, document summarization, coreference resolution. Probabilistic modeling is a core technique for many NLP tasks such as the ones

listed. In recent years, there has been increased interest in applying the benefits of Bayesian inference and nonparametric models to these problems.

One additional advantage of using probabilities in natural language processing is that it's possible to combine evidence from multiple sources in a systematic way. So for example, we can have a probability that counts from the speech recognition system. Then combine it with a probability from the text understanding system and so on and be able to build a better system that way. So the purpose of probability theory is to predict how likely it is that something is going to happen. One of the basic concepts in probability theory is the idea of an experiment or trial.

One important rule in probability theory is called the chain rule. It allows us to compute the so-called joint probability of multiple variables using a simple representation [3 :11]. So we want to compute the probability of *n* different events happening all at the same time. While this usually is very difficult because we have just many different combinations, so what we're going to do instead is apply the so-called chain rule, which works like this. If we have to compute the joint probability of n variables, we can just compute the probability of the first variable, so for example, w1. And then multiply this with the probability of the second variable given the first one, that's w2 given w1, times the probability of the third variable given both of the first two and so on, until the last term, which is the probability of wn, the last variable, given all of the previous ones. So this simplifies significantly the computation of the joint probability for P.

So this chain rule is used mainly in statistical and actual language processing, more specifically in Markov models, which is something we are going to talk about in the next lecture. So one more important property about probabilities is the idea of independence. So two events are independent if the product of their probabilities is equal to the probability of their intersection. So if unless a P(B) is equal to 0, this is equivalent to saying that the probability of A is equal to the conditional probability of A given B. So even if we have knowledge about the outcome of B, this is not going to affect our posterior understanding of the probability of A. This is going to be the same as the prior probability P(A). And just for completion here, if two events are not independent, we are going to call them dependent.

Removing constraints makes it less accurate to compute the probability. And also it makes it more statistical and feasible because there may be more instances in the trending data that have this particular combination of features. And one important observation here is that it's possible to do adding and removing constraints on the right hand side of the conditional probability.

The Bayes theorem, which is one of the most important topics in statistical natural language processing, based on the formula for joint probability. So, the joint probability of two variables, A and B, is equal to the probability of A, of one of them, times the probability of the other one, given the first one. So for example, what's the probability that today, I'm going to wake up late and then I'm going to take a walk. Well, that's the probability that today, I'm going to wake up late, times the probability that I'm going to take a walk. Given that I woke up late. So by symmetry, we can also add this as $p(A,B) = p(B)p(A|B)$. Now if you look at those two formulas,

you can see that the left-hand side is the same. And now we can therefore write this set of equations in a different format [1 :120]. So we can say that the conditional probability of B given A is equal to the conditional probability of A given B times the probability of B. Which comes from the second equation on the top, and then divided by the probability of A. So this equation here is the Bayes' Theorem, which is used everywhere in speech and language processing, also in computer vision, and statistics, and insurance companies, and finance and so on. So it's very useful because it allows us to compute the condition of probability of A given B, if we only know the condition of probability of B given A. Those two things are not the same. In fact they can be very different in some circumstances, so it's important to understand which one is which and also how to get from one to the next.

Now, one way to think about this formula in terms of statistical natural language processing, is to think of the problem of part of speech tagging. In part of speech tagging, you are given a word and you have to figure out if it's part of speech. For example, the word cat has to be labeled as a noun. So by looking at the word, cat, we may be able to predict with certain accuracy whether that word is a noun or not.

Conclusion. And of course, probabilistic NLP techniques meet the requirement for scalability [2: 10]. The execution time of the tagging process varies approximately linearly with the document side. Once the text has been tagged, retrieval and display tools are needed to allow the user to interact with the document. These use the tags to provide views on the document that reveal interesting properties and suppress the bulk of text. They do this in a way that is largely independent of the size of the document. Hence the user is protected from information overload by being able to be selective about the information they want to extract.

References

1. Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniquesm / D. Koller, N. Friedman. – L.: MIT Press, 2009. – 230 p.

2. Rush A., Sontag D. On Dual Decomposition and Linear Programming Relaxations for Natural Language Processing / A. Rush, D. Sontag // In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1–11. L.: Cambridge, MA, 2010. – 234 p.

3. Spitkovsky V., Hiyan A. Viterbi Training Improves Unsupervised Dependency Parsing / V. Spitkovsky, A. Hiyan // Proceedings of the Fourteenth Conference on Computational Natural Language Learning. – Uppsala: UYHJ, 2011. – p. 12-19.

4. Smith N. Linguistic Structure Prediction / N. Smith. – NY: Morgan&Claypool, 2011. – 353 p.