openMINTED
Open Mining Infrastructure for Text & Data

Facebook.com/openminted

Twitter.com/openminted_eu
#TDMstory

A COMPUTATIONAL LINGUIST TELLS HIS STORY

LINKING LANGUAGES WITH TDM

Dr Alan Akbik, Research Scientist at Zalando Research

HOW ARE YOU USING TDM?
My research focuses on the question of how to do TDM in the face of multilingual data. One big challenge is to create technologies that do TDM across different languages. A typical approach currently is to build a specialized system for each language. My research looks at how to do this automatically. If we have a technology that works well for English, can we simply transfer it to other languages such as German, Russian or even Chinese?

WHAT'S THE PRACTICAL APPLICATION?
In our search engines for fashion products, we have technologies that analyse the text that you've typed in for certain intents. These technologies have been engineered for one language. A lot of effort has gone into making them really good and with this type of research you try to bring similar technology to other languages. If Zalando decides to open a store in a Spanish speaking country, we don't want to build a search engine from scratch. We want to transfer the tools we've built for German to another language.

WHERE DO YOU GET DATA TO FUEL YOUR WORK?
One key source are corpora of humanly translated text data. A good example of this are European Parliament speeches. Every speech is officially translated into all European languages, and this data is made publicly available for research. If we have a tool that works well in English, we can apply it to the English portion of such a corpus, see what the technology finds and have the assumption that similar types of information will be contained in the translations. Using this data-driven approach, we can transfer the technology to another language.

IS COPYRIGHT AN ISSUE WHEN ACCESSING DATA?
The main issue really is clarity. We often discuss if we can just crawl the web to create very large corpora of linguistic data. There's a lot of uncertainty from our side in terms of to what extent that would be allowed. That's why we focus on established data sets but it would be a great boost to language understanding research to leverage the data in huge web corpora. If we knew specifically how far we are allowed to go when crawling data and using it for research, this would be very helpful.

HOW DO YOU SEE THE FUTURE OF TDM?
There are a few trends coming up where people will interact with machines in a more conversational way. People will type in fewer search queries and operate far more through colloquial voice commands. In the next 10 years I think we'll see models of language improving greatly. This will lead to a huge jump in language understanding.

LEARN MORE
https://research.zalando.com

European Commission

www.openminted.eu

OpenMinTeD is funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 654021.