Unlocking the Secrets of Language with **Inter-language Vector Space**

Andrzej Zydroń

Artificial intelligence (AI) is driving innovation and disruption in almost every industry. As companies are investing in solutions that will give them an advantage over competitors, plenty is at stake – AI solutions can boost productivity in a business setting by at least 40%. AI has become a well-used umbrella term, the reality is that it is multifaceted and multi-layered. In this piece I will speak about a new area of development in computational linguistic technology, specifically AI in the Natural Language Processing (NLP) field, we call it Inter-language Vector Space.

AI in localization

AI technology promises more efficient, streamlined and cost-effective workflows for those involved in the language business. Localization has not really benefited until relatively recently from AI. This technology has the potential to automate manual jobs across the localization supply chain, and an ongoing concern has been that this potentially can have a large impact on roles in the industry including those of linguists and project managers. The biggest impact so far beyond doubt has been on Machine Translation (MT), and specifically Neural Machine Translation (NMT). As the advances in NMT have reached a plateau in the last 3 years, the next big thing, furthered by the advancement in NLP, is the next generation automation, including Inter-language Vector Space.

Andrzej Zydroń is Chief Technical Officer at XTM International and technical architect of XTM Cloud, the enterprise cloud-based translation management system. He is one of the foremost IT experts on localization and open standards with over 30 years of experience in the industry. Andrzej sits and has sat on many open standard technical committees.
azydron@xtm.cloud

Vector Space: Next level of AI automation

Inter-language Vector Space is one of the big enablers for the next generation of automation. To fully appreciate why this statement may be true, we need to start with Vector Space. Vector Space came to the fore in 2013 with the publication of a seminal paper by the Google Research Center Team. Using Google's own vast news corpora, it was shown that using two algorithms simultaneously, and a vast neural network, you can predict the current word based on the context, and the surrounding words given the current word. This technology is able to work out relationships between words and how close their meanings are to one another. Each word is associated with a mathematical vector of 300 values which uniquely describes the word within the corpus and its relationship with other words that are of interest, a bit like a family tree. The resultant word-based data structures for the corpus are collectively called the Vector Space.

Now for the magic: Vector Space is able to work out detailed relationships between concepts that are truly amazing, such as if *king* is to *man* then what is the equivalent for *woman*, or if *Berlin* is the capital of *Germany*, then what is the equivalent for *France*, or if *Einstein* was a *scientist*, what was *Mozart*? It is also able to group similar concepts into clusters, such as *potato*, *salad*, *radish*, *broccoli*, *tomato* as belonging to one group while *apple*, *pear*, *orange*, *lemon*, *raspberry*, *blueberry*, *strawberry* as belonging to a different group, but both groups belonging to an edible plant group. Vector Space is also capable of working out semantic similarities between words such as adjectives and adverbs, e.g. *quick -> quickly*, *rapid -> rapidly*, etc., as well as opposites, such as *possible -> impossible*. This type of reasoning as you may deduct is fairly straightforward for the human brain.

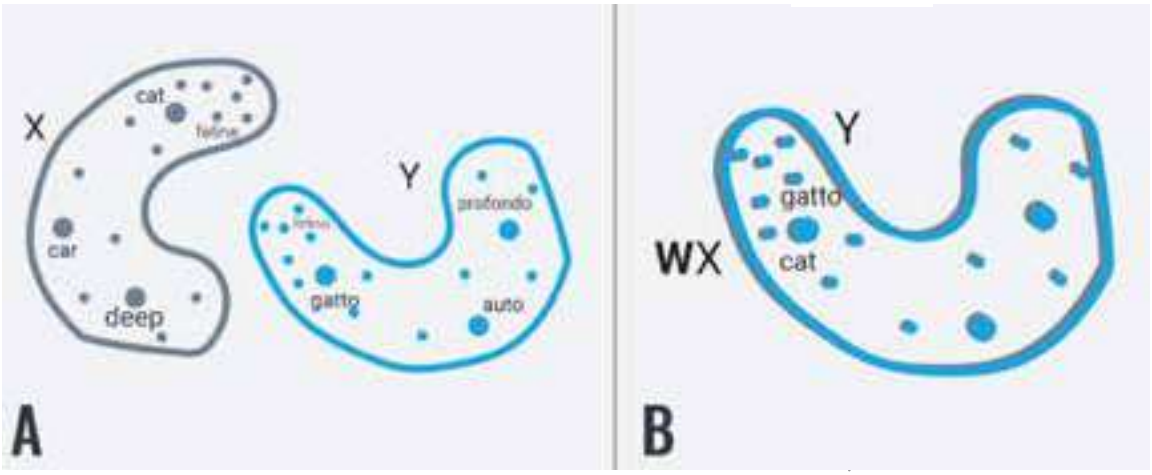
One of the key issues is the size of the corpus. Google's work was based on its own news corpus, plus Wikipedia in English only. Researchers at Facebook took up the challenge next and completed Vector Space data sets for 157 languages based on Wikipedia and a crawl of the complete Internet.

The Vector Space for a given language is unique to that language, and a limitation is that *you cannot compare entries between different languages*. This was a missing component in the work to date and this challenge was taken up by the XTM AI NLP Team which proved, based on work done by researchers at Babylon Health, that given appropriate bilingual data for two languages you can 'normalize' their Vector Spaces to create an Inter-language Vector Space. We can now,

XTM International

XTM International is a software development company that specializes in translation technology. It is the parent operator of the software package known as XTM Suite as well as XTM Cloud. The company is based in the United Kingdom, with offices in mainland Europe and a sales office in the United States.
https://xtm.cloud/

in addition to the semantic and syntactic features of Vector Space, add the probability of a given word in language A being a translation of a word in language B. We can also work out what words in language B are candidates for the translation of a given word in language A.



The overlapping of two Vector Spaces onto the same representation to make an aligned Inter-language Vector Space.

The essential part of the puzzle was producing a comprehensive and complete Vector Space for each language. One step was to compare the models based on Wikipedia with those based on a crawl of the whole of the Internet: it was plain that the Internet with the huge amount of text data in multiple languages won hands down – no corpus can compete with the comprehensive nature of a crawl of the complete Internet. It requires inordinately more processing power to calculate but it is able to provide a complete Vector Space model for a language. The XTM AI NLP team also has access to Big Data scale multilingual lexicons, with up to 15 million concepts per language, which produces a very high normalization factor for creating the Inter-language Vector Space. The results are remarkable.

Inter-language Vector Space vs. NMT – the difference

How does Inter-language Vector Space compare to Neural Machine Translation? Apart from the fact that both use enormous complex neural networks to achieve their aim, they serve different purposes. Neural Machine Translation is very much a black box in its operation. In goes a source segment and out comes the translation. You have no information regarding the individual words and phrases that make up the segment. Inter-language Vector Space allows you to look inside a translation, be it human or MT based and relate source and target words and phrases. Imagine you are travelling on a journey trying to get from A to B, NMT will map out a route, but will not tell you if there are road works or road closures on the way, that will make the

route impractical. It may take you down a single-track road or via a ford in a river. Inter-language Vector Space will inspect each part of the route to make sure that it is viable or not.

Typical applications of Vector Space in localization

Vector Space technology underpins functionality that enhances translators', reviewers' and correctors' productivity. The goal is to reduce human effort and speed up turnarounds so that all project participants can focus on more valuable tasks at hand. Here are some of the ways we have converted its power into technology features which simplifies and optimizes the translation process:

Automatic placement of inlines

Positioning inline elements is a chore that translators have to do when using a CAT tool for translation, thereby improving productivity and job satisfaction for the translators. When it comes to Machine Translation this is true for post-editors who can now rely on the automatic placement of inline elements rather than having to do this manually. Vector Space allows you to automatically position inline elements such as change of font markers, or hyperlinks, etc. This feature was released in XTM Cloud 12.3 in the XTM Workbench in April 2020.

Automatic corpus alignment

Vector Space allows for more accurate, automatic corpus alignment. The better, and more streamlined, the process is, the lower the costs and less human input are required. XTM's auto-align feature was enhanced with Inter-language Vector Space in XTM Cloud v12.4 (July 2020).

Bilingual terminology extraction

We all know too well that creating glossaries from scratch is difficult and very labor-intensive if it has to be done manually. The process itself is also not very rewarding for linguists as they have to align someone else's translation which they may not like very much. The Vector Space enabled functionality enables project managers to run bilingual terminology extraction during the alignment process to create glossaries faster. This feature was released in XTM Cloud 12.4.

Evolving Inter-language Vector Space

There are numerous functionalities that we have identified that this framework will enable, including various autocorrection systems, advanced predictive typing, verification and checking tools for NMT, and AI comparison engines.

Beyond that, Vector Space has countless other applications that we have not even thought of. It really does provide a 'Swiss Army knife' mechanism for increasing translation and post-edit productivity. Human beings will remain an important part of localization for years to come and Inter-language Vector Space will aid their productivity while helping to increase the quality of the output.

For the localization industry, AI opens up myriads of opportunities for growth and optimization of localization workflows. AI-driven automation, in particular, will continue to be an engine of innovation and source of competitive advantage during the current content economy.

As we can see, Inter-language Vector Space is an important additional component that can enhance and improve the output of both human and MT translation, both in terms of quality and productivity. In this respect, we believe it will be regarded as the most important advancement in translation technology since the advent of Neural MT. I feel that we are only just scraping the surface of the possibilities presented by this exciting new technology. As famed science writer and futurist Sir Arthur C. Clarke once put it, "Any sufficiently advanced technology is indistinguishable from magic" (1968), For us this technology is just that.