Type the words outdoor run into Google and you may be surprised to get more top results for guinea-pig pens than open-air jogging ideas. The English language constantly trips up search engines because so many of our words have multiple meanings. To add further complexity, language allows us to put words together into completely new sentences all the time. Current search engines work by simply matching key words: they have some understanding of what words mean, but not how they’re put together.

As humans, we understand each other because words and phrases are generally used in context, but computers just don’t have our world knowledge or any means of making judgements on what we’re really trying to say. Finding a way to help them overcome ambiguity and match our sophisticated understanding of language is a complex challenge. Solving it would enable vast improvements in the performance of language processing applications such as search engines and automated translation systems.

Stephen Clark is leading a large-scale research project that is taking a multi-disciplinary approach to the challenge. In collaboration with researchers at the Universities of Edinburgh, Oxford, Sussex and York he is developing a computational model that relies on mathematical representations of word meanings in a high-dimensional ‘information space’. The aim is to extend this idea to whole sentences, something that has never been considered before in computer science. The idea is that the meanings of phrases and sentences can then be compared automatically for similarity by observing how close the respective vectors are in the information space.

The work draws on the mathematics of quantum mechanics, which also uses linear algebra to model multi-dimensional spaces and suggest what happens when two quantum entities combine. Data from cognitive science is being used to test the model to ensure its results match with the answers humans would give. And by extending the model to compare similarities between images and extract meaning from them, this project is paving the way for a future in which computers can create better representations of the real world.

This work is funded by a three-year EPSRC grant and a five-year ERC grant, both from October 2012.

For more information see: https://sites.google.com/site/stephenclark609/research

Research themes:
» Human language technologies