Presentation: Linguamatics I2E and Machine Learning

Presenter: David Milward, CTO at Linguamatics.

“I'm going to talk about I2E and Machine Learning, and I'll start by talking about AI in general, NLP, and machine learning. Then talk about how I2E can be used for machine learning projects. Then talk about how machine learning can be used for I2E and within I2E. And then finally I'll talk a little bit about a learning NLP system.

So, artificial intelligence, I looked up the Wikipedia entry to make sure I was current on this, because I've been doing artificial intelligence for 30 years, or I would have expected it to be called artificial intelligence, but it does change a little bit. At times people tend to think of artificial intelligence for it to be intelligent it has to be something that is hard, therefore it's something that doesn't work. At various times, NLP has been regarded as part of AI, and certainly in the past when it didn't work, it was very much part of AI.

Once it started working, then people weren't quite so sure whether they wanted it to be classified as AI. Now that AI is fashionable again, NLP is definitely part of AI, and it always has been. It always has been.

So, I mentioned I'd say a little bit about machine learning, and so it's used in AI in general, but also as a technique with NLP. And again, machine learning methods have been very popular in NLP and probably for the last 20 years, and have really been regarded as one of the main paradigms in NLP.

Three main flavors, the supervised machine learning, where you use annotated data which maps between your inputs and your outputs, and you learn from that historic data where you're mapping from inputs to outputs. Semi-supervised, when you have machine analysis, but you have more of a human in the loop, so the human is saying whether something is good or bad more interactively. And then unsupervised, where you're trying to use large amounts of data and trying to understand patterns within that data.

There's been some recent successes with deep learning, and probably everyone here has heard about those. And those are based on neutral networks and used for both supervised and unsupervised machine learning. And some examples of things like machine translation using parallel corporate. And the deep learning has been able to generalize better than other techniques. And that means we can sometimes use a smaller data set than you would have previously done.

Things like image classification in medicine, there's been great strides there as well. But let's actually have a look at NLP also feeding into machine learning. So, if we take things like decision support, to actually do much of the decision support we need to access the knowledge from unstructured data. And so classically people say that 80% of the knowledge is unstructured. So, we need to get the NLP to get that information to then put into our machine learning models.

And a lot of AI projects are hampered because they only address projects where there's existing structured data. Worse, they'll use inappropriate structured data, for example ICD-9 codes, which are codes for billing. And they'll use those for research tasks where they're not really

appropriate. And we'll have a nice example of that later.
It's actually better if you can use the NLP to actually then get into the data and get the right answers from the data to then feed your machine learning models. We recently did a project to have a look at whether we could use the real health data records, so real health data are partnering with us and they have 8.9 million records. They're deidentified full text transcripts. And we had a look at whether we Linguamatics I2E and Machine Learning could predict the risks of opioid medication abuse. And Erin's done most of the work on this and will be presenting on this in November.

Up to 29% of patients are prescribed opioid drugs and then go onto misuse them. What we did was we identified various risk factors, things like alcohol abuse, illicit drug use, sexual abuse, psychiatric disorders. Those are then features which we can push into the machine learning model. We then selected cases where opioid misuse occurred, and we randomly selected equal sets of patients without misuse. And then trained the machine learning model, in this case a simple vector machine, an SDM model, to distinguish patients with misuse or not.

And so the results on the longitudinal data, we didn't have as much longitudinal data, so these results are probably ... the difference between these two results is more to do with quantity of data than actually the difference. So, the longitudinal data we were saying, okay, if we look in the historic data, can we then predict from the historic data whether someone's going to misuse the opioid drugs?

The general data didn't have the historic aspect of it. And really you do want the historic aspect ideally, but there was a problem with the amount of data where we had good longitudinal information. But even so, even with these figures of 71% recall, sensitivity, specificity of 79%, it's still getting pretty good percentages. I think rather better than we would have expected for this task.

That's an example that we did recently. I'll now go over to a customer case study. Chengyi Zheng presented this a few weeks ago as part of the Pistoia Webinar Series. This is a very nice use case where with gout flares, it's actually quite a difficult condition, because the population's generally old, and generally they have other conditions with overlapping symptoms. But you want to be able to actually identify people who do have a gout flare, because then the treatment is going to be different for a gout flare versus the other conditions.

The idea here was that I2E would concentrate on high recall, so returning patients with related signs or symptoms, various temporal aspects of their care, relationships to other diseases, implicit and explicit mentions of gout flare. Machine learning was then used to provide the precision. So, you're taking a set of patients who've got suitable signs and symptoms and then you're going to filter those down to a better set of patients using the machine learning.

I2E was used as the NLP core module, dealing with things like negation, dealing with pulling in the different ontologies, and producing a set of features. Those features are then going into the machine learning model, and then that's deciding whether the patient is likely to have a gout flareup or not.

And if we actually look at the results, the results, especially looking for the important case where people had greater or equal to three gout flares, the results here were very interesting, because in the two first bars on the left here, these were the two individual rheumatologists. And so their recall sensitivity was 74% and 83%. But actually when they got a consensus between them, then obviously it would have counted as 100%. but individually, they missed quite a few individual cases.

The actual system using I2E followed by machine learning actually did rather better. For sensitivity it got 93%. And this is actually the more important measure, because you really don't want to miss the patient's who've got the gout flares. If we compare that to the black bar, that's actually from the coding, the ICD-9 coding. So, using the structured codes actually was very poor for recall for this particular task. When it comes to specificity, it was doing slightly worse, but still very respectable, 84.6%. And that's actually a less important measure here, because you're not getting that many patients with gout flares, so you're prepared to go through some of the false positives.

I think this is one of the cases where it's actually a very interesting case to actually see where you can actually be doing slightly better than some of the actual clinicians at a task.

Okay, moving now, we've talked about how I2E can feed into machine learning, there's also machine learning within NLP and within I2E. Let's talk a little bit about some of the different flavors of that. Supervised machine learning, as we've said, that requires large scale representative annotated documents, and there are cases where you have that. So, for example for part of speech tagging there have been very good corporate built up over the years with part of speech tags. And most, if not all NLP systems use machine learning for their part of speech tagging, including I2E.

It's also used quite a lot for recognizing entities. So, if you're recognizing things like organizations, you look at the context around the organization to recognize that it's an organization. That's quite commonly used. Some people do plug that kind of approach into I2E, and I'll give an example of that later. Mostly with I2E people want to know which organization it is, which gene, which disease. They're slightly less interested in just it is a gene, or it is a disease. So, this kind of application has been less favored, but certainly it's a useful one, and some people will find it useful to plug this kind of approach in.

For extraction patterns, machine learning is used in the academic community. It's less commonly used for commercial systems, and that's because of the lack of annotated data. So, typically you have lots of different new tasks, and most of the issue is actually working out what the task really is. What the customer really wants to understand from their data. So, getting those requirements is actually an iterative approach, and something like I2E is very good at that.

If you ask people to annotate a large scale gold standard and spend several months at the beginning of a project, that takes far too long. And actually getting things like annotation guidelines to a good enough quality, you can end up with guidelines which are 50 pages long. So, that's a very big undertaking. For commercial systems, it's used rather less.

Semi-supervised machine learning is where you have more of a human in the loop. And we were actually as part of recent thinking about machine learning we were working out well okay, what are we doing in I2E? How does it relate to machine learning? How can we put more machine learning into I2E? And we realized that actually some of the systematic approaches that we're doing with the data driven approach, looking at frequency analysis, are equivalent to a semi-supervised machine learning approach. You're using the machine to give you suggestions, you're ranking them in frequency order to get the best possible suggestions, and then you're deciding which ones to keep and which ones to reject.

A lot of what we're doing is a semi-supervised machine learning approach. And that's useful when the tasks is initially ill defined, it puts the human in the loop to judge the suggestions. And it can provide good quality results very quickly. And that's been really useful for things like the feeding into machine learning. So, I should say that I just gave a couple of examples there, which are prominent recent ones, but people have been using I2E to feed into machine learning for about three years that I know of, possibly longer. But those are the ones I know about.

And I think part of the attraction is you can quite quickly get good quality features out. You can then test those features. Some of those aren't going to be useful for your model, some are. If they are useful for your model, you can then invest more effort into making those features really good.

Finally, I'll talk a little bit about unsupervised machine learning. The aim of that is you're taking a very large set of data and then you're creating something from that. So, the key example in NLP at the moment is what we call word embeddings. And the idea is that you take words and you learn the meaning of the word by the context it keeps, the other words around it. And people have had some nice examples of being able to learn what seems like a meaning. You can do this like England minus London equals France minus Paris and do some vector calculations and get a good correspondence.

So, you are getting some underlying notion of the meaning that way. And we've used it in the past with things like the words [inaudible 00:12:18] for finding synonyms. And we had a thesaurus explorer that we've also used for various custom projects. Using distributional similarity, and I talked about that a few years ago. And those have been reasonably used for finding similar words, but mostly similar words which are quite common words. Whereas typically I2E terminology discovery patterns are good for rarer words as well as some of the more common ones.

Typically we end up with a zip-slor effect. This zip-slor talks about the frequency of words, and you get this very long tail, a few frequent words and then a long tail. And you get a similar distribution whether it's for words or syntactic constructions. We want to automatically discover what's in the data using I2E, and we want to prioritize the most frequent constructions. But we also want to generalize to cover items in the tail.

To summarize, accessing text is key to the widespread use of AI and machine learning, and I2E's semi-supervised approach gives you an efficient way to capture the features in the text to then feed those models and get successful machine learning models which are using the bits of data you really want as opposed to using some of the structured data which isn't always appropriate.

Our current research is aiming to make the process of building the queries as systematic and efficient as possible. And so we're working on trying to incorporate more systematic processing, more machine learning within the query development process. And currently to I2E-51, you can already integrate with things like ontology editing, integrate with gold standard evaluation, and integrate with curation tools as part of learning NLP system. Thank you.”