Background: practical applications of corpus linguistics and text analytics

Professor Eric Atwell: This is Eric Atwell. I'm your lecturer for data mining and text analytics, and this is the first lecture of the module. So today I'm going to introduce you to your lecturer- that's me, Eric Atwell- give an overview of data mining text analytics module, tell you about the assessment for the module because that's very important for you, and then go into unit 1. The first part- background and practical applications of corpus linguistics and text analytics.

For this subunit you should be reading The language machine and also the assessment specification, both of which should be available in Minerva. So this is me. I'm Eric Atwell. I'm Professor of Artificial Intelligence for Language in the School of Computing at Leeds University. And there's my web page, or just Google Eric Atwell. I fixed it with Google so that I should be the first hit- the top most hit for Eric Atwell. My research is in Corpus linguistics and artificial intelligence for language, and in particular religious text analytics and Arabic and Islamic corpus linguistics, and also chat bots and AI for university education. And maybe we'll see some of this coming up in the course. And there's a picture of me, or rather an icon of me from the language machine book. I must be one of the few professors who has an icon specially drawn for them by the British Council.

OK so what's coming up in the overall module- well if you look at the Minerva page, you'll find more details- essentially like most modules, there are six units. And the first is an introduction to Corpus linguistics, text analytics, and sketch engine. And then unit two we'll look into more data about text data mining and a tool called Weka, Waikato environment for knowledge analysis. Everyone just calls it Weka. Unit three, we'll be looking at not just the text but the meanings of words and text because you want to compare two texts not in terms of what character strings they are but in terms of how they mean. Look at tagging of text and also for practical applications scaling to very large data sets. So what are the repercussions for classifiers if you're dealing with much bigger data sets.

Unit four, look at some examples of text analytics, in particular text analytics for machine translation and for information extraction. Having looked at Weka before and sketch engine as tool kits, you also look at some Python bits of program and tools for text analytics. Then we go on to a current research interest at Leeds University is chat bot and text analytics for University education, and such as teaching students. And then finally unit six, look at some current research in text analytics, and in particular BERT, a system developed by Google research labs. And there are

various versions of this coming out from other places too, but BERT is the main one, so we'll look at that.

OK. For reading, you'll find that quite a few of units have asked you to read chapters from this textbook Speech and Language Processing. Dan Jurafsky and James Martin have been working in this area for a long time, and they're now at the third edition of the textbook. This is very widely used for teaching text analytics, computational linguistics, natural language processing across the UK and around the world. So it's a good idea to use this book. Be aware that the current edition that you can buy online is only the second edition.

They're currently working on the third edition, which should be released by Pearson publishers. But luckily, Dan Jurafsky has a website at Stanford University where you can download the text of all the chapters of the third edition, of the draft text at least. And I'll be basing the course on this. And some of these chapters are core to the course, and I'll be asking you to read these. If you want to read more background, just read the whole book.

Another key textbook is on data mining by Ian Witten and his colleagues at Waikato University- Data Mining: Practical machine learning tools and techniques. This is not going to be dealt with in that much detail because we're going to using the Weka toolkit as a black box in essence. And if you want to find out more about the machine learning algorithms that are available, then there's a lot more details in the textbook. You can read this, but I'm not going to cover these on this course. There's a separate module on machine learning, another module on deep learning where you spend more time looking at the algorithms. But here, I'm going to be looking at how to use the algorithms to practical effect.

I'll also be asking you to read various research papers from conferences. A lot of research goes on in universities and in Google research labs, and so on. And when they come up with something interesting, it's generally published, publicized at a conference and then the papers appear in the conference proceedings, and they put on the web for everyone to read. And there's also some other material in websites I'd like to look at.

OK. You need to know how you're going to get marks for this. Well there's essentially two tests, an online test about halfway through the course to cover units one and two. And then right at the end of the course, there will be a test two, which will be twice as long and it will cover all the units from one to six. So 20% for the first test and 30% of your marks for the second test.

In between-- well, rather, over the course of the module, you'll be doing a sort of research project. And then at the end, you have to write up and submit a report of a max amount of seven pages long on this. And for the report, you'll be developing a research project proposal, which uses data mining and text analytics theory, methods, and technologies for some practical application of use

of your choice. So you have to decide what do you want to use data mining text analytics for, and then write a research proposal.

You could look at the Engineering and Physical Sciences Research Council guidance on how to write research project proposals, and there's a web link which should be in Minerva too. And I'll be telling you more about this later on. For now, you just have to think about what would be a useful application of data mining and text analytics, and how would you go about implementing that. So what's the objective of your research proposal? And what's the work plan for your research proposal?

You have to write well, six pages of English text. Obviously, you have to have a clear idea of what is the research hypothesis and objectives of what are you trying to achieve. And you have to have a program of methodology. How do you achieve this? What's the work? It could be over six months like an MSc project, or it could be over two or three years and like a bigger research project. That's up to you how long you take. But essentially, what are the work steps to achieve that? As well as the hypothesis and objectives and the work program, EPSRC also requires you to have some other parts.

So you have the background. What other work is going on that's like this? How is this a contribution to knowledge? What's novel about what you're doing? And how is it important? Is it going to make money for the UK or for your company or something? Well, how is it valued? As I say, we'll look into this in more detail later on. But these are the things that you have to do and to be marks for these things. And you also have to include a work plan diagram, such as a Gantt chart, which shows the stages in the work, and maps onto and visualizes the program that you've developed.

Notice in particular the work program should make use of appropriate methodology for AI projects. For example, CRISP-DM, which was mentioned- you should have covered in the data science module. That's the cross industry standard process for data mining. And it should also include at least two data mining or text analytics methods, techniques, or resources introduced in this module. So we're looking to see how have you applied various techniques or resources that you've learned about in the module in your research project.

And just by way of reminding you- so you might be asking- what's the difference between machine learning and data mining? So if you've done the machine learning course, or maybe you haven't yet but you will do sometime soon, this is mainly about learning the algorithms for doing various different sorts of machine learning, where you input some data and some classes if it's a classifier. And then you have various different ways of predicting the classes of unseen data. And you want to aim to have optimal accuracy, or to measure how good it is, how many of it you got right in your test set and how many are wrong.

For data mining, it's also machine learning, but it's applied and there's much more of a focus on using the machine learning as a sort of tool to tackle the practical problems. So what are the problems is more of an issue in data mining. Data mining is much more focused on data collection, data understanding, data and annotation, and data wrangling. If you look at Wikipedia, what is data wrangling, it's about taking data from various different sources and getting it into the right format so that your machine learning can actually work on it. Typically, you have to have something like a spreadsheet with rows and columns. Each row represents one instance, and each column represents some feature. So an instance is made up of a number of features. But the data in the first place isn't often like that. So you have to convert it into that form.

And Wikipedia will tell you data analysts typically spend the majority of their time in the process of data wrangling, compared to the actual analysis of the data. So getting hold of a data and getting it into the right format and understanding what is supposed to be is a big part of data mining. It's less so in machine learning typically for the machining exercises, you may have been given some data and just told to get on with it. So you don't really have much issue about getting hold of a data or deciding what data to use.

Notice in CRISP-DM the machine learning, the modeling bit, is only one of the phases. So here's the CRISP-DM phase is just by way of reminding you. And in data mining, it is very important- first of all- to understand what are the objectives and requirements that your customer or your business wants. And then how do you convert that into a data mining problem? Is it building a classifier? Or is it building some sort of clustering or something? So what is the machine learning you actually have to do? That's only part of the understanding of the problem.

And then you have to get hold of some data, collect the data from wherever it comes from. Try to figure out what it all means. What does this field mean or does that category mean? Check that it's OK at a defined data quality issues and if there's anything obvious you can straightforwardly see without doing any machine learning at all.

Then you have to then prepare it to get it into the right format, extract what records or what fields or attributes you ought to have. Maybe clean the data by throwing away rubbish you don't want. Finally, you do then get to do some machine learning, running the data through machine learning tools. But typically for data mining, you're not too bothered about implementing the learning algorithms. You probably just want to use some already implemented ones.

So you might just use a tool kit like Weka and click a button for SVN or click another button for perceptron. And then set the parameters by pull down menus, that sort of thing. Then the results come out and you have to evaluate not just in terms of which machine learning and which set of parameters gives you the most accuracy, but what meets the data mining objectives. And then

finally, if it's OK, putting the results into practice by some sort of repeated continuous mining of the data.

OK this is my way of reminding you should have seen this last in the data science module. OK so that's data mining. What about text analytics? What's text analytics? Well, it's data mining applied to text. So you might as well call it text mining, but text analytics seems to be a popular term in industry. Other terms abound as well. Computational linguistics is the part of linguistics or language science involving computation. Or within computer science and AI, it's often called natural language processing to differentiate from programming languages. It's natural languages. Jurafsky and Martin have decided to call it speech and language processing because speech is quite important. Language often is taken in text, and speech is quite different from text. Within linguistics, it's also called corpus linguistics. A corpus is a body of text, or a data set of text. So it's the linguistics to do with data sets, whereas computational linguistics is the linguistics to do with computation. But to do the computation, you have to have data sets. Or to analyze the corpus, you have to have computation. So actually corpus linguistics and computational linguistics are also very similar.

All of these things, cooperation linguistics, corpus linguistics, or natural language processing, they tend to focus on the theory and algorithms for doing things, whereas text analytics is the term preferred by industry. It's usually using computational linguistics or NLP as a sort of toolkit for tackling practical problems, making money in the end. Things like text data collection, understanding annotation, and wrangling are much more important in text analytics than they are in the theoretical fields of natural language processing.

If you look up in Wikipedia text analytics, you'll see your redirected to the text mining. So even Wikipedia thinks that text analytics is just a fancy word for text mining, or text data mining. All of this, you might think text is- it's English language is quite different from the sorts of things that are used to dealing with in machine learning- because for the machine learning course works, you probably had vectors of numbers. But if you can convert text into vectors of numbers, then all the algorithms that work for machine learning of images, or other data other numeric data also work for text. So the key thing in text analytics is having a way of taking the text, converting the words and sentences into numbers, or vectors of numbers. And then standard machine learning and data mining works just for that in the same way as for other types of data.

OK now I want to introduce this book that I want you to read. It's actually almost a magazine. It's not really a learned textbook. The British Council commissioned Eric Atwell, that's me, to write this mainly for the people working in British Council offices around the world. For those of you who don't know, Great Britain has embassies in most countries of the world, and the embassies look after passports and visas and stuff like that. They typically also have attached to them a British Council office. And a British Council promotes Great Britain, promotes British English, and English

language teaching, and so on. So the offices there wanted to know about computational linguistics. So the British Council decided it'd be nice to have a book about this.

A sort of magazine, if you look at it, you find there's diagrams and even jokes in it. It explores some of the technological, social, and educational implications of language machines in the years to come. You couldn't call it computational linguistics because that would be too complicated. So they called it the language machine. It's a survey of the current state of speech and language technology, highlighting history and academic disciplines contributing to the development of the technologies.

And also some of the practical problems and pitfalls, current and potential uses, predicted developments, and scenarios for the future. And notice this book is about 20 years old, but actually a lot of the concepts in there are still relevant. They still apply. And I think you'd be quite surprised how a 20-year-old book in IT is actually still quite relevant. It's partly because it's introductory, and the introductory part hasn't changed that much.

So one of the key things in text data analytics is to look at linguistics. So linguistics is the scientific study of language. At Leeds University, the linguistics department is part of the School of Languages, Cultures, and Societies, rather than being in the science faculty. But it is a sort of scientific approach. And it takes that language signal and analyzes it at various levels. So as phonetics, the study of speech, production, and perception. Lexicography, or Lexis, is a study of words or vocabulary items in a dictionary for example. And the words will have the way that they written, the way that they're spoken, the meaning, and the big grammatical function. And then the syntax or grammar, the study of grammatical arrangements of words and morphemes. And morpheme is a minimal unit. We'll see more of this later on.

Semantics is the study of meaning, the meanings of words, the meaning of sentences, and how these are related to each other. How can you tell that marriage and wedding mean the same thing, even though they're different character sequences? Pragmatics is the analysis of language in practical use. Take into account the context of use. Discourse modeling is about phenomena that range over more than one utterance in a discourse or dialogue. So in a chat bot, I say something. Then the chat bot says something, and then I say something. Then the chatbot says something. These are not independent somethings. They come in a pattern of some sort.

So the book also- the language machine text- also considers what's the point? Why do people want to be interested in the language machine or text analytics? Well as academics, we're interested in computer models of language. And it's also good to have language resources, like a corpus, or text data set, or a dictionary. These are useful for machine learning. They're also useful for people. Text analytics is also useful as a way of communication between people and

computers. You don't want to have to type in numbers. Maybe you want to be able to speak to your computers and listen to your computer. It's also between people, it can assist communication.

For example, when I go to China to teach, it's helpful that we have machine translation so that I don't have to learn Chinese. Or it's very good for monitoring communication between people. Social media is basically people posting to other people. And the tech scientists can analyze their social media, see what themes are coming up, stop posts which are fake news, and all sorts of useful stuff like that. And of course, another big reason for developing tech analytics is there's money in it. So the government and the industry are interested in wealth creation.

OK. And the book also looks at what the challenges are for text analytics. Well, one big problem for the UK and for British companies is it's very expensive because the BIG IT companies like Google, Apple, Amazon, Microsoft, IBM, Facebook, they have huge resources, really big research labs. I, as an individual researcher with a dozen PhD students, I can't really compete with Google research labs in terms of their abilities to do things. They can develop very general text analytics tools. I can hope to develop very specialist niche tools, like of the Quran. I don't really think Google is not that interested in that. That's the sort of thing that we can still do.

Another problem is it's quite difficult to elicit user requirements. If you ask people, what would you like from your text analytics, since they don't know what text analytics can do, they don't really know what they want from it. And they do sort of expect it will be able to speak proper English, natural English. That means that if and when the speech recognition system gets it wrong, or the speech synthesis system says something silly, then they just give up. People tend to not tolerate bad performance, or poor performance, or less than perfect performance from text analytics. And there are some tasks where it's just not appropriate.

For example, inputting data into a spreadsheet or understanding the data in a spreadsheet, it doesn't really make sense to try and do this in simple English. This is just numbers of the things in tables. You can't convert a table into English sentences very straightforwardly. And we also- the way that we interact with computers uses devices like keyboards and touch screens- and if we're going to start using English language instead, then we have to think about that. A touch screen becomes less useful if I'm talking to a computer. Also the people, the users, need training and time to learn these new interactive methods. If I'm used to using a keyboard, I'm now supposed to dictate my essays rather than type them. It's quite difficult to say things in a way which is natural and correct first time. And many applications have all of these problems to them.

OK. The language machine also looks, in some detail, UK and European Union research initiatives in text analytics. This is relevant to your coursework exercise because you have to come up with a research proposal. Bear in mind that these are- first of all these are quite large scale research projects- and they're only talking about mainly the book gives some of the titles and topics without

going to the details of the actual research plan. Also note that this is 20 years old, so some of the research projects are a bit dated perhaps. For example, developing machine translation.

Now we have Google Translate, it's not such a big thing to do anymore. And the UK government has said, or the engineering and physical science Research Council rather has said, it's probably not worth funding research into speech interfaces in the UK anymore because we just can't compete with the likes of Amazon Echo and other. Google and so on have already got this sorted. OK.

The European Union tends to fund research projects requiring several partners. So that's not just a project for one person or one group, but for several. For example, I'm involved here at Leeds in the EduBot project with three other universities and two companies over two years. And if you want to find out more, we'll look at this later on in the course. One interesting thing in the language machine is it has a snapshot of what's called the BT technology calendar. Back in the 1990s, BT decided it would be useful to actually pay people to try to predict what was coming in the future in terms of technology. And there's some very wacky ideas, some good ideas, but it turned out not all of them work very well. So here's just an example.

They thought by the year 2003, IT literacy would be essential for any employment. But even now it is still possible to get some jobs where you don't have to be IT literate. You probably have to be able to type in a document to get your CV, but there are jobs around where it's not necessary. By 2005, they thought there'd be full voice interaction with machines. Well, OK, you can talk to computers for some purposes. But most of the time when you use the computer, you're actually probably using a touch screen or a keyboard. You're not talking to it. So that didn't really take off.

They said by 2007, there'd be domestic robots, small attractive robots everywhere in all homes. Well, OK, some homes have got a robot vacuum cleaner, but it tends to be not very good. I mean, it doesn't actually pick up much dust. It's not very intelligent. It's not very functional.

So again, robots did not take off. And this is partly because domestic duties are- they thought that the easy things could be done by robots- was actually it's very complicated to clean a house or to cook a meal. And it's far too complicated for robots to learn. Robots can do things like building cars. I used to work in a car factory making Vauxhall cars, and that job is now gone. The robots have taken over. In fact, if by 2012 robots will be available for almost any job at home or in hospitals. Again hospital cleaners and hospital nurses, again that hasn't happened. Those are actually for AI, the difficult things. They may be semi-skilled or not valued for human labor, but it's far too difficult for robots to do this.

They thought by 2018 AI would be imitating thinking process of the brain. Well, we know what that meant. So we don't know if that's true or not. But certainly by 2025, they thought that thought recognition would be useful for input output. You wouldn't need to type, or use a touch screen, or

even talk to the computer. You simply think and it would work. They even thought that you wouldn't have to learn anything. You wouldn't have to go to University to learn stuff. You simply plug-in and somehow we're learning via the thought interface. So you wouldn't even have to plug it in. You just have to think about things, and you know them somehow. Again, I don't think it's going to happen by 2025. They certainly thought by 2030, human brain intelligence would be enhanced by just linking it to AI. Maybe. I don't know.

The book finishes with some examples of real applications. For example, soldiers in Bosnia and that shows you the years in the 1990s, the war in Bosnia was a big thing. There were British soldiers and other soldiers from other countries there. The soldiers in Bosnia wear a small computer on their chests, and say to it hands up, or get out of the car, or other things that soldiers have cause to order Bosnian civilians to do. And it would say it in Bosnian, or Croat, or whatever it was acquired. Another thing they predicted was text editing, smart tools to check grammar, idioms, and style. Well, you've got these in Microsoft Word.

Another thing they predicted or actually saw was an application that was coming. AltaVista. AltaVista at the time was the big web search company owned by the computer giant digital. Digital was the second largest computer company in the world after IBM. And they had a huge buildings in Leeds. They had a huge center just outside the White Rose shopping center as it is now. That's gone now. Digital disappeared because it decided to focus on mainframe computers, rather than these newfangled personal computers and handheld devices.

Anyway, AltaVista launched a free machine translation service on the internet, so you could get a web page translated from French into English for free. Good idea, isn't it? And they also had Lufthansa, the German airline, had ALF, a friendly flight information service, which can hold conversations with callers at 300 airports. So you could phone in and say, when is my plane from Berlin to Dresden. And it would tell you, I suppose. Another thing is that car and lorry drivers can use voice commands to activate the phone, and also to listen to email messages, and dictate replies to emails on the move. Well this is a nice example of technology which was technically possible, but it turned out to be not very safe to do. So nowadays, you're not supposed to listen to email messages or dictate replies while you're driving because you're being distracted from driving. OK then.

So in summary, I have introduced you to myself. I'm Eric Atwell. Google Eric Atwell and find out more about me. If you go to my webpage, you'll hear some happy songs you can listen to. You have an overview of the data mining and text analytics module and the assessments that are coming. There'll be a couple of online tests at the beginning and at the end. But most importantly, I'm going to be asking you individually to come up with a data mining and text analytics research project proposal, and write it up in the style of an EPSRC research grant proposal following the guidelines there.

OK. So for unit one, the first part, we've looked at some practical applications of corpus linguistics and text analytics from the language machine. This book, I'd like you to read it in some more detail and also read the assessment specification in more detail. And hopefully you'll feel comfortable with that. And then later on in the module, I'll give you some more details as to what you actually have to do to make sure you all come up with really good research proposals. OK. Thank you for watching and listening, and I hope you enjoy the module. Bye for now.

[END]