LIE : Leadership, Influence and Expertise

R. Catizone*, L. Guthrie*, A.J. Thomas*, and Y. Wilks****

*Oxford Internet Institute

** Florida Institute of Human and Machine Cognition

*** Oxford University Press

E-mail: louiseguth@gmail.com, arthur.thomas@oii-ox.ac.uk, ywilks@ihmc.us, roberta.catizone@oup.com

Abstract

This paper describes our research into methods for inferring social and instrumental roles and relationships from document and discourse corpora. The goal is to identify the roles of initial authors and participants in internet discussions with respect to leadership, influence and expertise. Web documents, forums and blogs provide data from which the relationships between these concepts are empirically derived and compared. Using techniques from Natural Language Processing (NLP), characterizations of authority and expertise are hypothesized and then tested to see if these pick out the same or different participants as may be chosen by techniques based on social network analysis (Huffaker 2010) see if they pick out the same discourse participants for any given level of these qualities (i.e. leadership, expertise and influence). Our methods could be applied, in principle, to any domain topic, but this paper will describe an initial investigation into two subject areas where a range of differing opinions are available and which differ in the nature of their appeals to authority and truth: 'genetic engineering' and a 'Muslim Forum'. The available online corpora for these topics contain discussions from a variety of users with different levels of expertise, backgrounds and personalities.

Keywords: Expertise-detection, Authority-detection, Reputation-power

1. Introduction

This paper describes our research into methods for inferring social and instrumental roles and relationships within document and discourse corpora. The goal is to identify the roles of participants in internet discussions with respect to leadership, influence and expertise, concepts which are closely related but not identical. For example, Einstein in his early years had expertise but not influence, whereas Nostradamus had a great deal of influence but, by our standards, no expertise. Some people have both and most people have neither. Our main goal will be to identify these features of postings and to determine what, if any, the relationship is between expertise, on the one hand, and notions of authority, influence and leadership on the other, where we shall take those last terms as broadly synonymous. Our research is at an early stage, although we shall describe here some promising initial results.

We have three ways in which to explore this relationship. First, using web documents, forums and blogs to provide data from which the relationships between these concepts can be empirically derived and compared, we want to use techniques from Natural Language Processing (NLP) as characterizations of authority and expertise to see if these pick out the same or different participants as may be chosen by techniques based on social network analysis (e.g. Huffaker, 2010, Joinson et al., 2010). NLP methods for exploring these concepts have already been pioneered by e.g. (Strzalkowski et al., 2011 and Bender et al., 2011). We shall describe the social science methods as *external* to the texts, in that they are concerned mostly with relationships between texts considered as closed entities, e.g. the frequency of posting and which posters initiate threads. We shall describe NLP methods, by contrast, as methods *internal* to the content of postings and derived from the content of the texts themselves. Later, we shall explore the possibility of deriving stronger measures of our concepts by combining the internal and external measures if they do in fact pick out the same concepts.

Secondly, although our methods could be applied, in principle, to any domain topic, this paper will describe an initial investigation into two subject areas where a range of differing opinions are available and which differ in the nature of their appeals to authority and truth: 'genetic engineering' and a 'Muslim forum'. The available online materials for these topics contain discussions from a variety of users with different levels of expertise, backgrounds, personalities and motivations. The corpora collected in the first phase of this project are described below. One area we are investigating is whether measures of expertise derived from the scientific area, where there are more objective measures of truth and expertise, can be transferred to the other area where these notions are more difficult to pin down.

Thirdly, we want to throw some light on the complex issue of finding some "gold standard" verification of the machine-based annotations that result from these investigations and which might be the basis for further machine learning and training. The Muslim corpus is internally marked by participants with a "reputation power" score (https://www.vbulletin.com/docs/html/vboptions_reputation). We are investigating whether the reputation power is in fact indicative of either of our target concepts, although our aim here is not to reverse engineer the algorithm that constructs the reputation power scores, but only to see if we can use it as a step to assigning our target concepts to postings reliably.

2. Inferring Levels of Expertise

This investigation began by identifying:

1) the levels of expertise of participants in a discourse or information-seeking activity,

2) the ways in which those levels may appear to change as the conversation or investigation evolves in time, and

3) what those relative levels of expertise may allow us to infer about the relative roles of the participants within the posted discussion.

The focus of the work is on relatively informal communications such as blogs and forums, looking both at how discourse evolves in time, as well as across multiple blogs and forums in space (where a single participant may use a different *nom de plume* for different postings, but where it may be possible to cross-identify him or her on the basis of level of expertise and style). While we are aware of the dangers of restricting attention artificially to particular knowledge domains, and the difficulty of generalising our approach across domains, we believe that useful progress can be made by choosing one or two domains of independent interest, discourse about which is characterized by a wide range of expertise and possible intentions (including sometimes the intention to mislead or obfuscate). The two domains which we are investigating as rich sources of material include:

(a) A Muslim forum as noted above, where a wide variety of opinions are expressed very robustly, and where there tend to be leaders or 'advice givers' who have more expertise and subject knowledge, in contrast to followers who are often seeking advice. In this domain, we are attempting to characterize leadership, expertise, influence and relationships with a "bottom up" approach based on NLP or internal techniques, and

(b) the domain of *genetic engineering*, specifically, the various internet newsgroups on "Do-It-Yourself biology" (a new and rapidly developing domain where scientific "amateurs" – in the best sense of that word -- discuss how to perform experiments at home which until recently would have been possible only in university labs where conversations and information seeking can be characterized fairly precisely by assessing the level of expertise of a conversational player with respect to an existing body of scientific knowledge, as represented by articles on corresponding subjects from peer-reviewed scientific journals. With our NLP-based approach we aim to identify those participants who have expertise, or are seeking to develop it (the "serious players"), and to distinguish them from participants whose interest may be more amateurish, however robustly expressed, and therefore probably of less significance.

Both of these domains provide a large body of publicly available and relevant textual materials, from formal sources (e.g. scientific publications) and informal ones (e.g. forums, blogs, newsgroups), which ease the task of creating training corpora and identifying and comparing language models. The difficulty in both cases is to establish suitable criteria against which to judge our assessments of expertise and leadership.

Machine learning and statistical techniques are used to derive language models that should serve to identify levels of expertise of individual participants in conversations about the domains, by using established closeness measures between corpora so as to assess the distance of a given text from an established body of attested expert text in the same domain. Methods of proven utility include Latent Semantic Analysis (Landauer and Dumais 2008) and related probabilistic approaches such as Latent Dirichlet Allocation (Blei *et al.* 2003), which allow the extraction of language models specific to appearance of certain clusters of "concepts" in a text, as well as the classification of texts on the basis of occurrences of terms related to those concepts. These levels of expertise will be calibrated with respect to recognized benchmarks (e.g. text from a Letter to *Nature* will carry a greater presumption of expertise than a posting on a student blog), and appropriate confidence metrics for these are being developed. In the case of the Muslim forum, we have attempted to find analogues of scientific expertise by taking comparable corpora from both the Qur'an itself, together with scholarly comment on it, and from a set of well-formed articles in Islamic topics (e.g. "Islam and marriage," Islam and children") on Wikipedia. We have also examined the value of combinations of more syntactic metrics that we believe may indicate expertise (and initial results confirm this, see below) such as word length, sentence length, vocabulary size and readability measures.

3. Experiments on expertise assessment

3.1 Readability

One first hypothesis was to see if readability is correlated with expertise. There are many standard measures of the former (such as flesch, ari, coleman and smog) and for the latter we took, as possible initial measures, the reputation power of postings in the case of the Muslim forum, and the source of scientific documents (as between high-grade journals and DIY biology forums) in the case of genetics. A sample of many such results is shown below in table 1 for the grade level required to read the genetics documents on four such measures, where the leftmost four columns are journals and the two rightmost are forums. In the cases of all four systems, the forums have a significantly lower grade level needed to read them and, for at least on the first three algorithms, this feature would suffice to pick out the greater expertise needed to read the journals as opposed to the forums. Similar results were obtained for measures of reading-ease (from fleschease at least), percentages of long and short sentences, words of more than six letters, words of at least three syllables, and percentage vocabulary overlap with the top 1K words from a large gigaword corpus of English (e.g. in the last case a significantly higher proportion of words came from the commonest 1k words in blogs than in science journal papers). One possible problem with this approach is that the readability measures may be a matter of <u>style</u> rather than expertise, in that the same author (with a single given expertise level) might write quite differently in a forum and in a journal paper.

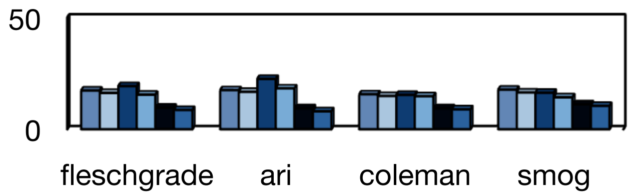

Table 1: School grade level needed to read scientific articles and forums

For the Muslim texts, the readability scores make less significant distinctions between those postings ranked with high and low reputations powers However, the proportions of words deemed "positive" or "strong" seems inversely correlated with posts with the highest reputation power. In table 2 below, the columns represent deciles of reputation-power increasing to the right (and the two leftmost columns should be ignored as the postings with zero reputation power implies unranked rather than being ranked zero). The categories are from the General Inquirer dictionary: (http://www.wjh.harvard.edu/~inquirer/homecat.htm).

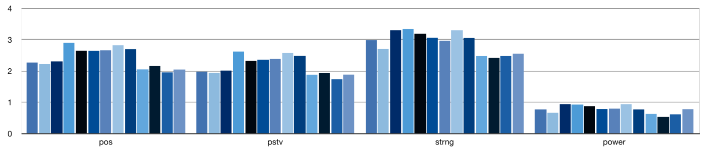

Table 2: "strong" words in Muslim forum posts of differing reputation powers

In summary, some stylistic measures do pick out higher expertise in both science and the Muslim forum (in so far as expertise in the latter can be expressed as higher reputation power, an assumption we shall question later), but using totally different measures in the two domains.

3.2 Vector methods

We mentioned above a range of standard comparison models between texts deriving from Information Retrieval which model the "closeness" between texts in terms of a word vector model, and these are often referred to as "geometric methods". We trained the Latent Semantic Analysis (LSA/Gensim (Rehurek and Sojka, 2010)), Latent Dirichlet Allocation (LDA/Gensim), Random Projections (RP/Gensim) and Naïve Bayes (NB/Mallet) tools on a number of different biology and Ummah corpora and subcorpora.

The overall goal of these classifiers is to support the ranking of postings and/or posters according to their "similarity" (in terms of their use of similar "concepts" or "topics") to those which might be accepted as "expert" or "authoritative" such as:
in biology, the US National Library of Medicine "Medline" collections of abstracts of journal articles on selected topics which are also discussed in the DIY biology forums;
in the Muslim forum collection, posts by contributors who have been assigned high (>25) reputation power, who are the most prolific, or most highly quoted by others . We also used the collection of articles from Wikipedia on Islamic subjects, and an English translation of the Qur'an and its associated scholarly footnotes (Rodwell, 1909) as proxies for expertise.

The kinds of questions we hope to be able to answer using these methods include:
How similar is the collection of postings by a given contributor to the postings on similar topics in the "expert" or "authoritative" collections?
Are postings by the most prolific, most quoted or highest reputation contributors more similar to the expert collections than the average in those forums?
What are the topics of different threads, and can we relate different threads by the commonality of their topics? This may allow us to understand how threads evolve, split, etc.
In the Muslim corpus, is there a correlation between expertise metrics, reputation scores, and the frequency of use of (actual or transliterated) Arabic words?

In order to validate the classifiers, we wrote software to split a corpus randomly into a labelled training and test sets. Classifiers can then be validated by comparing the classification predicted by the classifier against the assigned labels. Some details of the corpora on which we have trained can be seen in table 3 below.

The main feature of interest emerged from comparing posters of apparently low expertise with the authoritative corpora, namely significant journals in science and the Qur'an itself or Wikipedia Islamic topics in the case of the Muslim forum, and where apparently low expertise was identified with posting in the DIY blogs or having reputation power of <25 respectively. In both domains the LSI algorithm identified individual posters with low reputation but whose posts had a significantly higher similarity score to the authoritative texts than the lower reputation posters in general. We could think of these as "hidden experts" whose reputation should be higher based on their similarity rankings, or again one could take this, in the case of Ummah, as further criticism of the reputation-power measure as an indicator of expertise. The Ummah posters were categorised in a number of ways on "external" criteria so as to how candidates for expertise: the most frequent posters, the posters who started threads, the posters most quoted by other, following, posters and so on. In the case of each category "hidden experts" emerged under the definition above, in some cases the same poster in several of the categories. In the case of the biology forums, it was clear that some posters wrote texts far closer to the journal articles than others and it would be a reasonable inference that they are in fact experts contributing to forums. An interesting result in the case of similarity between Ummah postings and the Qur'an itself is that posters with reputation power <25 were significantly closer to the Qur'an itself than those with reputation power >25 which suggests that a style closer to the Qur'an does *not* gain an author reputation power and one might speculate that such authors are seen as less original. Alternatively, again,

“Authoritative” Corpus	Content	No. of docs
muslim_general_rep_GT_25	all posts by authors with reputation power > 25	5,610
muslim_general_top100_posters	all posts by authors amongst the top 100 most frequent posters	13,777
muslim_general_top100_thread_starters	all posts by authors amongst the top 100 topic initiators	9,201
muslim_general_top_quoted	all posts by authors who have been quoted by others	4,706
Koran_Rodwell	Rodwell’s translation of the Qur’an and all notes	116
Wikipedia_Islamic_Topics	set of Wikipedia entries on a range of Islamic topics	83
anthrax_merged	selected abstracts from Medline relating to genetic engineering of anthrax bacteria	7,246
molecularbio	postings from the Internetgroup “molecularbio” relating to genetic engineering of bacteria	1,904
bionet_molbio_methods_reagents	postings from the “BioNet” interest group on methods and reagents for molecular biology	3,327

Table 3: Details of corpora used for training

this may be evidence that reputation power is not an indicator of expertise *if* closeness to the Qur’an were to define expertise instead of reputation-power.

4. Determination of authority and influence

Attempts have been made to set out taxonomies of relationships among discourse participants, usually expressed in terms of intention types (e.g., *seeking information*, *receiving assistance*, *becoming influential*, etc.), as well as relationships based on these types (e.g., *colleague, subordinate, teacher, student*, *influencer*, etc. Chulef *et al*. (2001) proposed a taxonomy of intentions that largely apply to very general notions such as *happiness* and *health*. One hypothesis of this work is that the structure of the discourse, together with information about pair-wise relationships of individuals, if available, will lead to more accurate detection of intentions. Resources used for this part of the work include:

(1) tools for identification of dialogue acts from informal communications:
(e.g. Khosravi and Wilks 1999; Webb, Hepple and Wilks 2008).

(2) tools for detection of social roles/relationships using control,disagreement, and involvement (e.g.Strzalkowski *et al*. 2011).

We tried two initial “internal” measures as candidates for capturing a notion of authority or leadership. We first applied Webb’s (Webb et al., 2008) Dialogue Act tagger to the Muslim corpus, as a whole and in subsets, so as to locate those acts that could plausibly be associated with authority, such as forms of urging others to act at the behest of the writer. The algorithm was run without further training, in the GATE platform (Cunningham et al., 1996), as experience has shown that extensive training to new corpora does not produce much variation. The first result is that nearly every post (about 94% of sentences) is tagged as some form of statement. We derived from other Dialogue Act categories an initial set of patterns -–augmented by inspection and hand-coding---with which to identify postings that might be deemed AUTHORITY-DIRECTION. This derivation is dependent on some of the smaller categories detected by the Webb tagger, and also contains some 30 two or three word patterns such as “check out,” ”go for it”, “don’t encourage,” and “don’t influence”. This is simply a seed set from which we expect to bootstrap a larger set if the postings located by these strings are of interest.

The average "reputation power" of the whole corpus is 15.4 and the average "reputation power" of the set of sentences identified by this pattern set is just 6.8. This is an interesting and unexpected result which could be interpreted in two quite different ways:
a) the “authority” sentences have low reputation power
b) the reputation power measure is flawed because these sentences might be independently expected to be more significant.

We cannot decide between these possibilities until we are able to construct, from a wide range of features, an alternative “prestige” measure independent of reputation power but some function of the features we are determining, by NLP and social measures.

Secondly, we investigated how far the number of transliterated Arabic words in a posting compares with reputation power. We first created a detector for transliterated Arabic words in Roman letters by seeking distinctive trigrams in a large transliterated Arabic corpus (the Buckwalter corpus: http://www.qamus.org/wordlist.htm) that are not present in an English corpus. This algorithm then picked out Arabic transliterations in the Muslim corpus with a high level of precision, although without extensive annotation we were not able to test its recall. There were 44,725 Arabic words detected in the entire corpus (of 1.5 million words). The average number of Arabic words per post is 2.4, but the average reputation power of the posts which have more than the average number of Arabic words is 12.1 which is still lower than the average reputation power of all posts (15.4). This method may be contrasted with that of (Marin et al., 2010) who sought authority claims in Wikipedia discussion pages. In their domain, they were able to use reference to Wikipedia policy documents as the key feature that marked out authority claiming sentences, which is precisely the "parallel " authority we do not have here. However, when we have the most plausible grown set of text markers for this feature we may attempt to use Koranic invocations or even (non-formulaic) Arabic words as offering a similar external validating authority feature in the Muslim corpus.

5. Conclusion

Our initial results suggest some simple "internal" methods for detecting expertise in both the scientific and social texts, but further work will be needed to show if the methods derived for the former can be applied to the latter. Some initial attempts to define authority in political texts seem promising but throw considerable doubt on the value of reputation-power as any kind of "gold standard" or proxy for this feature of text. Further work will be needed to seek correlations between the internal NLP measures applied here and those deriving from "external", social science, measures.

6. Acknowledgements

This work was supported by a grant from the UK Government.

7. References

Bender, E., Morgan, J., Oxley, M., Zachry, M., Hutchinson, B. , Marin, A., Zhang, B., and Ostendorf, M. (2011). Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. *Proceedings of the ACL Workshop on Language in Social Media*, pp. 48–57.

D. Blei, A. Ng, and M. Jordan. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3:* pp. 993–1022.

Chulef, A., Read, S., and Walsh, D. (2001) A hierarchical taxonomy of human goals. *Motivation and Emotion. 25:3.*

Cunningham, H., Wilks, Y., Gaizauskas, R. (1996) GATE -- a General Architecture for Text Engineering. *In Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*, Copenhagen.

993–1022..

Huffaker, D.A. (2010). Dimensions of Leadership and Social Influence in Online Communities. *Human Communication Research,* 36(4), pp. 593–617.

Joinson, A.N., Reips, U-D., Buchanan, T.B., and Paine Schofield, C.B. (2010). Privacy, Trust and Self-Disclosure Online. *Human-Computer Interaction*, 25, pp.1 – 24.

Khosravi, H., and Wilks, Y. (1999). Routing email by function not topic. *Journal of Natural Language Engineering.* vol. 5 (3).

Landauer, T. and Dumais, S. (2008*), Scholarpedia, 3(11):*4356. doi:10.4249/scholarpedia.4356

Rehurek, R. and Sojka, P. (2010) Software Framework for Topic Modelling with Large Corpora, *Proc. LREC 2011*

Marin, A., Ostendorf, M., Zhang, B., Morgan, J. T., Oxley, M., Zachry, M., Bender, E. M. (2010) *Detecting Authority Bids in Online Discussions*. SLT 2010: IEEE Workshop on Spoken Language Technology, December 12-15, 2010, Berkeley,CA.

Rodwell, J. M. (1909) *The Koran* (Everyman Library, J.M. Dent and Sons, London 1909) as downloaded from Project Gutenberg at http://www.gutenberg.org/ebooks/2800

Strzalkowski,T.,Broadwell, G., Stromer-Galley, J., Shaikh, S., Liu, T., and Taylor, S. (2011). Modeling Socio-Cultural Phenomena in Online Multi-Party Discourse. *Analyzing Microtext: Papers from the 2011 AAAI Workshop* (WS-11-05)

Webb, N.,Liu, T.,Hepple, M., Wilks, Y. (2008). Cross Domain Dialogue Act Tagging. In *Proc. of the 6th International Conference on Language, Resources and Evaluation (LREC'08)*, Marrakech, Morocco.