Discourse Analytics

Carolyn Penstein Rosé, *Language Technologies Institute and Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213*

Abstract. This chapter introduces the challenging and multi-disciplinary area of Discourse Analytics. Discourse analytics has its impact in multiple areas, including offering analytic lenses to support research, enabling formative and summative assessment, enabling of dynamic and context sensitive triggering of interventions to improve the effectiveness of learning activities, and provision of reflection tools such as reports and feedback after learning activities in support of both learning and instruction. It draws from the full gamut of modeling technologies including supervised, semi-supervised, and unsupervised modeling approaches. The purpose of this chapter is to encourage both an appropriate level of hope and an appropriate level of skepticism for what is possible in the area of Discourse Analytics while also exposing the reader to the breadth of expertise that one needs access to in order to do meaningful work in this area. It is not the goal to impart the needed expertise – that would be beyond the scope of a short chapter like this. Instead, the goal is for the reader to find his or her place within this scope so that it is possible to know what kinds of collaborators to seek in order to form a team that encompasses sufficient breadth. It would be difficult indeed for one researcher to fully possess all of the expertise needed. We begin with a definition of the field, casting a broad net both theoretically and methodologically. We explore both representational and algorithmic dimensions of this area. We conclude with suggestions for next steps for readers who are interested in delving deeper into this area.

DEFINING THE AREA

Discourse Analytics is one area within the field of Learning Analytics (Buckingham-Shum, 2013; Buckingham-Shum, de Laat, de Liddo, Ferguson, & Whitelock, 2014). It includes processing of open response questions in educational contexts, and a large proportion of research in the area focuses on assessment of writing, but it encompasses more than that, including analysis of discussions occurring in discussion forums, chat rooms, microblogs, blogs, and even wikis. We consider Learning Analytics broadly as learning about learning by listening to learners learn, with our listening normally assisted by data mining and machine learning technologies, though the published work in the area may precede but not yet include automation in all cases (Knight & Littleton, 2015; Milligan, 2015). Furthermore, we consider that what makes this area distinct is that the listening focuses on natural language data in all of the streams in which that data is produced, such as answers, instructional explanations including self-explanations, think aloud utterances, and discussion within pairs, small groups, and even communities.

This chapter offers a very brief introduction to this area situated within the field of Learning Analytics broadly. Discourse Analytics is an area that has alternately suffered from two dangerous misconceptions. The first is an extreme over-expectation fueled by the desire of many to have an off-the-shelf solution that will do their analysis work for them at the click of a button. Those falling prey to this misconception are almost certainly doomed to disappointment. Making effective use of either the most simple or the most powerful modeling technologies requires a lot of preparation, effort, and expertise. The second is an extreme skepticism, sometimes resulting from disappointments arising from starting with the first misconception, or other times coming from a deep enough understanding of the complexities of discourse that it is difficult to get past the understanding that no computer could ever fully grasp the nuances that are there. While it is true that discourse is incredibly complex, it is still true that there are meaningful patterns that state-of-the-art modeling approaches are able to identify. Much published work from recent Learning Analytics and Knowledge and related conferences that illustrate the state-of-the-art are cited throughout this chapter. A recent survey on computational sociolinguistics tells the story from the perspective of the field of Language Technologies (Nguyen, Dogruoz, Rosé & de Jong, in press), and might be of interest to dedicated readers.

A short chapter such as this one cannot possibly do justice to the whole area of Discourse Analytics. The hope of this chapter is that it provides helpful pointers to readers who want to dig a little

further. Two previous workshops on the topic of Discourse Analytics survey foundational work within the Learning Analytics community (Buckingham-Shum, 2013; Buckingham-Shum, de Laat, de Liddo, Ferguson, & Whitelock, 2014). An extensive overview of issues and methods situated more narrowly within the field of Computer-Supported Collaborative Learning can be found in three earlier published journal articles (Rosé et al., 2008; Mu et al., 2012; Gweon et al., 2013). A short course in the area can be found in the Text mining unit of the Fall 2014 Data, Analytics, and Learning[1] MOOC offered on the edX platform. Other resources will be presented at the end of this chapter.

In this chapter, we are interested in the natural language uttered during episodes of learning. We seek to be theoretically and methodologically inclusive. Much of the existing work on discourse analytics views learning and its connection with language from a Cognitive lens, in other words, seeking categories of language behavior whose presence in a discourse makes predictions about learning gains because of the connection between the associated discourse processes and cognitive processes associated with learning. In this chapter, we seek to view learning and its connection with language through a Social lens in order to leverage the important interplay between Cognitive and Social factors in learning (Hmelo-Silver et al., 2013; O'Donnell & King, 1999). For example, we seek to identify discourse processes that reveal underlying dispositions, attitudes, and relationships that play a supporting (or sometimes interfering) role in the learning interactions. Regardless of the situation in which it is uttered, natural language is deeply personal and deeply cultural. Embedded within it are artifacts of our personal experiences and those of generations that came before us. The details of language choices provide clues about the identities we purposefully project as well as sometimes those we seek to hide or even those we are not consciously aware of. They project assumptions about and attitudes towards our audience and our positioning with respect to our audience, or sometimes just assumptions we want our audience to think we are making. We use these choices as currency in an economy of relationships in which we seek to achieve goals that we have adopted (Ribeiro, 2006).

With this understanding, as we use computation as a lens to aid in our listening to learners, we must acknowledge that we are always abdicating some of the responsibility for interpretation to the technologies that sit in between us and the learning processes, including whatever was lost or transformed in the recording into some digital form, and the further reduction and transformation that occurred during the application of the analytic technology (Morrow & Brown, 1994). With that caveat in mind, in this chapter we will focus heavily on questions of model interpretation and assessment of validity.

SCOPE AND FOCUS OF THIS CHAPTER

When one initially thinks of analytics, it is common to immediately think about algorithms (Witten, Frank, and Hall, 2011). However, it is important to take a lesson from applied statistics and instead think about representation first. At the heart of Discourse Analytics work is a focus on representation of the data. Machine learning models cannot be applied directly to texts. Rather, the predictor features must be extracted from the text. These predictor features can be conceived as questions of the form "Is __ found in the text?" or "How many times is __ found in the text?". If each feature is one of these questions, then for each instance, the feature value is the answer to the question. Interested readers can get a good feel for the breadth of simple features that can readily be extracted from text and what impact they have on predictive accuracy of classification models by experimenting with the publically available LightSIDE tool bench[2] (Mayfield & Rosé, 2013), a freely available, off-the-shelf workbench with an extensive user's manual, example data sets, instruction about process, and contact information for researchers who are willing to offer help.

The key to success with modeling technologies applied to text is to ask the right questions, which produce clues that are meaningful. Thinking about this question begins by considering how language is structured. Though on the surface language may appear to the naked eye as a monolithic, unstructured whole, the fact is that it is composed of multiple layers of structure, each described within a separate area

[1] https://www.edx.org/course/data-analytics-learning-utarlingtonx-link5-10x
[2] http://lightsidelabs.com/research/

of linguistics. An introductory survey of linguistics textbook (O'Grady, Archibald, Aronoff& Rees-Miller, 2009) would be a valuable resource for researchers desiring to get into this area of Learning Analytics. At the finest grained structure is the sound structure level, referred to as phonology and phonetics. Here the basic sound units of a language and how they fit together into the syllabic structure of a language are described. A basic alphabet of sounds comprise the set of phonemes, but within dialects these may be pronounced in particular ways, which carry social significance because of their association with a host of socially relevant variables such as ethnicity, socio-economic status, and region. Just above that level, the inner structure of words is described in a layer referred to as morphology. This is where systems of affixes we learn in our grammar classes come into the picture, which change the tenses on verbs or number on nouns, among other things. Above that is the level of syntax, where the grammatical structure of whole sentences is described. Also at the level of a sentence is the area of semantics, which describes how meaning is composed through fixed expressions, by convention, or by composing smaller units, guided through syntax, and referencing low level semantic units at the level of lexical semantics. Above the sentence level is the level of discourse, where we find rhetorical strategies among other aspects of structure. While these technical terms might be unfamiliar to many readers, they may provide useful search terms for readers who desire to find relevant resources for further reading.

If one traces the history of several areas in which natural language data has been the target of automated analysis, we hear the same refrain, namely the key to valid modelling is design of meaningful representations. The hope in including this example in this chapter is that readers can be spared from learning the same lesson the hard way. Taking an example focusing on one of the earliest applications in the area of Discourse Analytics, one of the earliest cases where this lesson was well learned was that of automated essay scoring (Page, 1966; Shermis & Hammer, 2012). The earliest approaches used simple models, like regression, and simple features, such as counting average sentence length, number of long words, and length of essay. These approaches were highly successful in terms of reliability of assignment of numeric scores (e.g., Shermis & Burstein, 2003), however they were criticized for lack of validity in their usage of evidence for assessment. In later work the focus shifted to identification of features more like what instructors included in their own rubrics for scoring writing. This investigation led to inclusion of content focused features, including techniques akin to factor analysis such as Latent Semantic Analysis (Foltz, 1996) or Latent Dirichlet Allocation (Blei et al., 2003; Griffiths & Steyvers, 2004) to aid in content based assessments, though these still fall prey to problems with unigram features since they are also usually grounded in a unigram language representation. Other factor analytic language analysis approaches such as CohMetrix (McNamara & Graesser, 2012) have recently been used for assessment of student writing along multiple dimensions, including such factors as cognitive complexity. In highly causal domains, that build in some level of syntactic structural analysis have shown benefits (Rosé & VanLehn, 2005). In science education, success with assessment of open ended responses has been achieved with LightSIDE (Nehm et al., 2012; Mayfield & Rosé, 2013), the freely available suite of software tools mentioned earlier in this chapter, which supports use of text mining technology by non-experts, which includes a wide range of sublexical, lexical, syntactic, and template based feature extractors (Gianfortoni et al., 2011).

At this point it is useful to return to the tension between over and under-expectation of Discourse Analytics. If we think about the challenges in identifying appropriate, meaningful features, we must come to terms with the limitations of the lenses we construct through modeling tools. The analytic technologies applied in Discourse Analytics may serve as a lens in the hands of researchers or practitioners that sits in between them and the episodes of learning that occur within the world, or they may be a filter that mediates the interaction between learners and instructors, between learners, or between learners and learning technologies. Lenses are useful precisely because they do not simply transfer the exact details of the world viewed through them. Instead they accentuate aspects of those images that would not as effectively been seen without them. That's what we need them to do. At the same time they obscure other details that are deemed less interesting by design. Lenses always distort. But in order to use them in a valid way, we must understand what each accentuates and obscures so that we can select an appropriate lens, and so we can interpret what we see in a valid way, always questioning

how the picture would be different without it or with a different lens. Thus, from the beginning, we would caution those who consume the research in this area, develop these lenses, or actively apply them in research or practice, to be wary of what is inevitably lost or transformed in the process of application. Now this chapter will turn its attention to specific areas within the scope of Discourse Analytics.

REPRESENTATION OF TEXT

Key decisions are made at the representation stage that strongly influence how the data will appear through the analytic lens. At the representation stage, text is transformed from a seemingly monolithic whole to a set of features that are said to be extracted from it. Each feature extractor asks a question of the text, and the answer that the text gives is the value of the corresponding feature within the representation. Imagine that all you knew about a person was the set of answers to questions posed during a game of twenty questions, and now your task is to classify that person into a number of social categories that are of interest. If the questions are carefully constructed, you may be able to make an accurate prediction, but nevertheless, you must acknowledge that much information and insight into that person as an individual will have been lost in the process. Once information is lost at this important stage in the process, it cannot be recovered through application of an algorithm, no matter how advanced and generally effective that algorithm is. Thus, we emphasize throughout this chapter the importance of careful decision making about representation, careful reflection about interpretation, and careful questioning of the validity of inferences made. While readers new to this area may find these caveats somewhat illusive, they will become more clear with experience.

Overview

The most typical kind of feature extractors used in text mining problems are what are called unigram features. In the case of a unigram feature space, for each word appearing within the set of texts in the training data, there will be a corresponding feature that asks about the presence of that word within each text. While unigram feature spaces frequently achieve reasonably high performance, the models often fail to generalize beyond data collected under very similar circumstances to that of the training data. The reason for the lack of generalization is that these unigram models essentially memorize for each class value label in a very surface fashion what kinds of things people talk about in the set of instances associated with that label in the training data. If there is some consistency in that, then it can be learned by these models, but that consistency rarely generalizes very far. Generalization comes when the features that are extracted come from a relevant layer of structure.

The purpose of the feature based representation of text is frequently to enable predictive modeling for classification or numerical assessment, where the objective is to achieve this predictive modeling with the highest possible accuracy (Rosé et al., 2008; McLaren et al., 2007; Allen, Snow, McNamera, 2015). This orientation will be the focus of this section. However, it is important to note that in some work within the broad area of Discourse Analytics, the representation work is the focus, and meaning is made of the identified predictive features, and thus the predictive modeling if any serves mainly as a validation of the meaningfulness of the identified features (Simsek, Sandor, Buckingham-Shum, 2015 ;Dascalu, Dessus, McNamera, 2015; Snow, Allen, Jacovina, Perret, McNamera, 2015).

With respect to predictive modeling for classification, in this vector based comparison, the chosen features should make instances that are of different categories look far apart within the vector space, and instances that are of the same category look close within the vector space. This principle can be used also to trouble shoot a text representation. Features that either make instances that should be classified the same way look different or make instances that should be classified differently look similar are very likely to cause confusions in the classifications made by models trained using representations that include those features. The problem is often either ambiguous features (i.e., features that mean different things in different contexts, but the representation does not enable leveraging that context in order to disambiguate) or fragmentation (i.e., the same abstract feature is being represented by several more specific features, some of which are missing or too sparse in your data). It may also be that the most meaningful features

are simply missing from your feature space, and other features, which may correlate with the meaningful ones within the specific data used as training data, will often "steal the weight", which ends up being counter-productive when the model is applied to new data where the spurious correlations between the meaningful features and less meaningful features may not exist or may be different.

Case Study

In order to illustrate the thinking that goes into representation of text for Discourse Analytics, we will start with a common example, namely analysis of affect in text, otherwise known as *sentiment analysis* (Pang & Lee, 2008). It is one of the most heavily marketed applications of text mining, and it is frequently the first thing researchers think to apply to their text data when they are faced with analysing it. We will begin by introducing some issues in this area of text analytics and conclude with an investigation of what these analytics do or do not offer in terms of explaining patterns of attrition in MOOCs, where one might reasonably expect to see more expressions of negative affect from students who are struggling and ultimately drop out. We will see that the picture is far more complex than that (Wen et al., 2014a). In leading the reader through this case study, the hope is that the reader will see how one might progress through cycles of data analysis from pre-conceptions that start out overly-simplistic, but become more informed through iteration. The most interesting work in the area of Discourse Analytics, or any are of analytics applied to rich, relatively unstructured data, will follow a similar storyline.

Simplistic treatments of sentiment identify texts as exhibiting either a positive or negative sentiment, and rely on an association between words and this affective judgment. Thus, much work has gone into the construction of sentiment lexicons, which associate words with a positivity or negativity score. The area of sentiment analysis is well developed, and has gained a substantial representation in industry, providing services to businesses related to marketing issues. Nevertheless, the limitations of the technology are clear.

An important consideration is that the text by itself is not enough to fully gauge the level of positivity or negativity. Demographic and situational variables may need to be taken into account as well. For example, it is known that prevalence of and selection of swearing words depends upon age, gender, socioeconomic status, and characteristics of the addressee (McEnery, 2006). Similarly, the context of speech determines the level of formality assumed, which also has an impact (Biber & Conrad, 2011).

Furthermore, what is learned from examination of the linguistic literature is that much about attitude is not conveyed in text through words that are specifically positive or negative (Martin & White, 2005). This can be illustrated with the following example related to the weather. A statement such as "The weather is beautiful today." contains the required positive word, however, "The sun is shining." is only obviously positive if one knows that typically sunny days are preferred over rainy days. "It's a great day for staying indoors." indicates that the weather is not so good, despite the presence of a positive word. "My rain boots are feeling neglected." could easily be taken as a positive comment about the weather despite the presence of a negative word. "Snowflakes that dance past my window in formation grab my attention." gives the sense of a more positive attitude than "A flurry of snow is the first thing I notice as I check the weather through the window." though neither contain any specifically positive or negative associated words and could arguably be used to describe the same scene.

The limitations of sentiment analysis technology are less obvious under certain circumstances. For example, on texts that are specifically aimed at the purpose of expressing an attitude (such as movie reviews and product reviews) and that are long enough that the person writing eventually states their attitude directly, despite the large amounts of nuance that will be dismissed, performance of the technology is at its best since these texts are most in line with the assumptions behind the approach. Even in these contexts, there are a range of easy and hard examples. For example, "simplistic , silly and tedious" is obviously negative, but this one is less so: "While the ensemble player who gained notice in guy ritchie's lock , stock and two smoking barrels and snatch has the bod , he's unlikely to become a household name on the basis of his first starring vehicle".

Now we will investigate situations more close to home where the approach may fall short. Because sentiment analysis is one of the most widely known and widely used language technologies by

researchers and practitioners in other fields who are interested in text, it is not surprising that analysis of forum data from MOOCs is one area where we find applications of this technology, and thus that work will be a convenient case study. The rationale for its application was that discussion forum data may be useful for understanding better how, why, and when students drop out of MOOCs, with the idea that students may drop out because they are dissatisfied with a course, and that dissatisfaction should be visible using sentiment analysis as a lens. In an early such investigation, however, Ramesh et al. (2003) found no relation between overall sentiment expressed by students (as assessed using a totally automated method) and their associated probability of course completion. Adamopolous (2013) developed a sentiment related assessment method to measure sentiment associated with different course affordances in order to understand what students express their attitudes about in course discussion forums. They used a combination of automatically identified sentiment expressions paired with a grounded theory approach to identify themes in the course aspects mentioned in connection with attitudes. With this more detailed view, they were able to identify that not attitude in general, but attitude towards Professor, Assignments, and other Course materials had the strongest association with dropout. Other our later work (Wen et al., 2014a) we pushed the automated analysis further, increasing the accuracy of sentiment measurement, and contrasting sentiment expressed by a student vs sentiment they were exposed to as well as contrasting sentiment at the student level with sentiment at the course level. In this work, the exact connection between sentiment related variables and dropout depended upon the nature of the course.

With more probing, it became clear that a far more nuanced way of characterizing affect in posts was needed. For example, negative affect expressed in purely social exchanges might be disclosure leading to enhanced emotional connection. Problem talk in a problem solving course might just indicate engagement with the material. Negative affect words, expressions, and images may come up in a literature course where stories about unfortunate or stressful events are discussed, and yet that expressed sentiment might have nothing to do with a student's feeling about the experience of reading that material or even discussing that material. We conclude that sentiment analysis is not as simple as counting positive and negative words. Individual words are not enough evidence of attitude, context matters. Some rhetorical strategies combine negative and positive comments in the same review, and sometimes sentiment is expressed indirectly. Nuances like this observed through qualitative analysis must be taken into account when representing your data.

UNSUPERVISED METHODS

A variety of factor analytic (Garson, 2013; Loehlin, 2004) and latent variable analysis techniques (Skrondal & Rabe-Hesketh, 2003; Collins & Lanza, 2010) have been popular in the area. These may be unsupervised (i.e., not requiring pre-assigned labels), supervised (i.e., requiring examples to have pre-defined labels), or lightly supervised (i.e., requiring some external guidance to learning algorithms, but not requiring a pre-assigned label for every example). In this section, we focus on unsupervised methods. The most popular such techniques in the education space include factor analytics approaches like Latent Semantic Analysis (Foltz, 1996) or structured latent variable models like Latent Dirichlet Allocation (Blei et al., 2003) mentioned briefly above. Thus, here we delve slightly deeper into the details and discuss strengths and limitations. In recent work in Learning Analytics, unsupervised approaches have been used for exploratory data analysis (Jolsimovic et al., 2015; Sekiya, Marsuda & Yamaguchi, 2015; Chen, Chen, & Xing, 2015), sometimes paired with visualization techniques (Hsiao & Awasthi, 2015), or alternating with or building on hand analysis (Molenaar & Chiu, 2015; Ezan-Can et al., 2015). These modelling technologies have widely been used because researchers think of them as approximating an analysis of textual meaning. The reality is that they are much less apt at doing so than the prevailing view would have one believe. These tools do indeed have their place in the arsenal of Discourse Analytics tools. However, the hope of this chapter is to raise the curiosity of the reader to dig a little deeper in order to foster an appropriate scepticism as described above.

Topic modeling approaches have become very popular for modeling a variety of characteristics of unlabeled data. A well known and widely used approach is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is a generative model and is effective for uncovering the thematic structure of a document

collection. Hidden markov modelling and other sequence modelling approaches are becoming popular for capturing progressions in student experiences (Molenaar & Chiu, 2015). Sometimes these approaches are combined in order to identify how language expression changes in predictable ways over times in terms of the representations of thematic content (Jo & Rosé, 2015). Statistical approaches such as these are meant to capture regularities. They are most valuable as tools in methodologies that value data reduction and simplification. Because they dismiss as noise the unusual occurrences within the data, they are less valuable in methodologies that seek unusual happenings that challenge assumptions. Though one might adopt an anomaly detection approach to identify instances that violate assumptions as a way of identifying such examples, in practice the examples found are more likely to be unusual in ways that are not necessarily interesting from the standpoint of challenging assumptions of theoretical import.

LDA works by associating words together within a latent word class that frequently occur together within the same document. The learned structure is more complex than traditional latent class models, where the latent structure is a probabilistic assignment of each whole data point (which is a document) to a single latent class (Collins and Lanza, 2010). An additional layer of structure is included in an LDA model such that words within documents are probabilistically assigned to latent classes in such a way that data points can be viewed as mixtures of latent classes. This structure is important for topic analysis. By allowing the representation of documents as arbitrary mixtures of latent word classes, it is possible then to keep the number of latent classes down to a manageable size while still capturing the flexible way themes can be blended within individual documents. Each latent word class is represented as a distribution of words. The words that rank most highly in the distribution are the words that are treated as most characteristic of the associated latent class, or topic.

Because LDA is an unsupervised language processing technique, it would not be reasonable to expect that the identified themes would exactly match human intuition about organization of topic themes, and yet as a technique that models word co-occurrence associations, it can be expected to identify some things that would be expected to be associated. At heart, LDA is a data reduction technique. Its strengths lie in identification of word associations that are very common in a corpus, which frequently correspond to common themes. However, the common themes do not necessarily have a one-to-one correspondence with the themes of interest. Unfortunately, that means within the resulting representation, there will not be a distinct representation for those themes of interest that are not common. Similarly, unusual phrasings of common ideas will also typically fail to map to an intuitive representation within the LDA space. Representation of the textual data is also an important consideration. Typically LDA models are computed over feature spaces composed of individual word features. Thus, whatever is not captured by individual words will not be accessible to the model.

In light of these caveats, readers who are interested in fostering their own tangible sense of appropriate scepticism can try this exercise: Let's assume the researcher has begun with a research question and some theoretical framing in which the investigation is being conducted. For each topic then, the researcher might seek to identify some theoretical constructs within the framework suggested by the content displayed within the top ranking texts as associated with that topic. The association between that topic and the construct can then be subjected to a face validity check by sorting all documents by the topic association, and then checking to see whether the construct is strongly represented at the top and absent at the bottom, with some middling association in the middle. Dimensions within the model that strongly match human intuition when subjected to such as face validity check can be marked with that intuitive label, and others can be treated as meaningless dimensions that "soak up" the words that are not strongly thematic within the corpus. When applying the model to some downstream task, such as associating comments with concepts in a course, one must carefully consider what important associations will be missed simply because they are not within the scope of what the model is able to capture.

SUPERVISED METHODS

At the other end of the spectrum are supervised methods. Taking a somewhat overly simplistic view, supervised machine learning methods are typically algorithms that operate over sets of vectors that associate a collection of predictor features, often referred to as attributes, with an outcome feature, often

referred to as a class value. Recently, applications of supervised machine learning have been applied to the problem of assessment of learning processes in discussion. This problem is referred to as automatic collaborative-learning process analysis. Automatic analysis of collaborative processes has value for real time assessment during collaborative learning, for dynamically triggering supportive interventions in the midst of collaborative-learning sessions, and for facilitating efficient analysis of collaborative-learning processes at a grand scale. This dynamic approach has been demonstrated to be more effective than an otherwise equivalent static approach to support (Kumar et al., 2007). Early work in automated collaborative learning process analysis focused on text-based interactions and click stream data (Soller & Lesgold, 2000; Erkens & Janssen, 2008; Rosé et al., 2008; McLaren. Scheuer, de Laat, Hever, de Groot & Rosé, 2007; Mu, Stegman, Mayfield, Rosé, & Fischer, 2012). Early work towards analysis of collaborative processes from speech has begun to emerge as well (Gweon et al., 2013; Gweon, Agarwal, Udani, Raj, & Rosé, 2011). A consistent finding is that representations motivated by theoretical frameworks from linguistics and psychology show particular promise (Rosé & Tovares, in press; Wen et al., 2014b; Gweon et al., 2013; Rosé & VanLehn, 2005). We have already mentioned the LightSIDE toolbench as a good place to start getting experience in this area.

MOVING AHEAD

Readers who are interested in getting more familiar with the area of Discourse Analytics would benefit from digging first into some foundational literature. It is grounded in the fields of linguistics (Levinson, 1983; O'Grady & Archibald, 2009), Discourse Analysis (Martin & Rose, 2003; Martin & White, 2005; Biber & Conrad, 2011), and language technologies (Manning & Schuetze, 1999; Jurafsky & Martin, 2009; Jackson & Moulinier, 2002). Happy reading!

References

Adamopoulos, P. (2003). What makes a great mooc? an interdisciplinary analysis of student retention in online courses. In Proceedings of the 34th International Conference on Information Systems, ICIS.

Allen, L., Snow, E., & McNamera, D. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques, in Proceedings of Learning Analytics and Knowledge 2015.

Biber, D. & Conrad, S. (2011). Register, Genre, and Style, Cambridge University Press.

Blei, D., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation. J. Mach. Learn. Res (3) 993-1022.

Buckingham-Shum, S. (2013). Proceedings of the 1st International Workshop on Discourse-Centric Analytics, workshop held in conjunction with Learning, Analytics and Knowledge 2013.

Buckingham-Shum, S., de Laat, M., de Liddo, A., Ferguson, R., & Whitelock, D. (2014). Proceedings of the 2nd International Workshop on Discourse-Centric Analytics, workshop held in conjunction with Learning, Analytics and Knowledge 2014.

Chen, B., Chen, X., & Xing, W. (2015). "Twitter Archaeology" of Learning Analytics and Knowledge Conferences, in Proceedings of Learning Analytics and Knowledge 2015.

Collins, L. & Lanza, S. T. (2010). Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral, and Health Sciences, Wiley.

Dascalu, M., Dessus, P., McNamera, D. (2015). Discourse cohesion: A signature of collaboration, in Proceedings of Learning Analytics and Knowledge 2015.

de Wever, B., Schellens T., Valcke, M., & Van Keer H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. Computers and Education, 46, 6-28.

Erkens G, Janssen J. (2008). Automatic coding of dialogue acts in collaboration protocols. Int J Computr Support Collab Learn 3:447–470.

Ezan-Can, A., Boyer, K., Kellog, S., Booth, S. (2015). Unsupervised Modeling for Understanding MOOC Discussion Forums: A Learning Analytics Approach, in Proceedings of Learning Analytics and Knowledge 2015.

Foltz, P. (1996). Latent Semantic Analysis for text-based research, Behavior Research Methods, Instruments, & Computers 28(2), pp197-202.

Garson, G. D. (2013). Factor Analysis, Statistical Associates Publishing

Gianfortoni, P., Adamson, D. & Rosé, C. P. (2011). Modeling Stylistic Variation in Social Media with Stretchy Patterns, in Proceedings of First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties, Edinburgh, Scottland, UK, pp 49-59.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101,5228-5235.

Gweon, G., Jain, M., Mc Donough, J., Raj, B., Rosé, C. P. (2013). Measuring Prevalence of Other-Oriented Transactive Contributions Using an Automated Measure of Speech Style Accommodation, International Journal of Computer Supported Collaborative Learning 8(2), pp 245-265.

Gweon G, Agarwal P, Udani M, Raj B, Rosé CP. (2011). The automatic assessment of knowledge integration processes in project teams, In: Proceedings of the 9th International Computer Supported Collaborative Learning Conference, Volume 1: Long Papers, 462–469.

Hmelo-Silver, C., Chinn, C., Chan, C., & O'Donnell, A. (2013). The International Handbook of Collaborative Learning, Routledge.

Hsiao, I and Awasthi (2015). Topic Facet Modeling: Semantic and Visual Analytics for Online Discussion Forums, in Proceedings of Learning Analytics and Knowledge 2015.

Jackson, P. & Moulinier, I. (2007). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization, John Benjamins Publishing Company.

Jo, Y. & Rosé, C. P. (2015). Time Series Analysis of Nursing Notes for Mortality Prediction via State Transition Topic Models, in Proceedings of The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)

Jolsimovic, S., Kovanovic, V., Jonanovic, J., Zouaq, A., Gaesevic, D., Hatala, M. (2015). What do cMOOC participants talk about in Social Media? A Topic Analysis of Discourse in a cMOOC, in Proceedings of Learning Analytics and Knowledge 2015.

Jurafsky, D. & Martin, J. (2009). Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, Pearson.

Knight, S. & Littleton, K. (2015). Developing a Multiple-Document-Processing Performance Assessment for Epistemic Literacy, in Proceedings of Learning Analytics and Knowledge 2015.

Kumar R, Rosé CP, Wang YC, Joshi M, Robinson A. Tutorial dialogue as adaptive collaborative learning support, In: Proceedings of the 2007 conference on Artifcial Intelligence in Education: Building Technology Rich Learning Contexts That Work, 2007, 383–390.

Levinson, S. (1983). Pragmatics (Chapter 6, Conversational Structure), Cambridge Textbook in Linguistics

Loehlin, J. C. (2004). Latent Variable Models: an introduction to factor, path, and structural equation analysis, Routledge.

Manning, C. & Schuetze, H. (1999). Foundations of Statistical Natural Language Processing, The MIT Press.

Martin, J. & Rose, D. (2003). Working with Discourse: Meaning Beyond the Clause, Continuum

Martin, J. & White, P. (2005). The Language of Evaluation: Appraisal in English, Palgrave

Mayfeld E, Rosé CP. (2013). LightSIDE: open source machine learning for text accessible to non-experts. In: Handbook of Automated Essay Grading. Routledge Academic Press.

McEnery, T. (2006). Swearing in English: Bad language, purity, and power from 1586 to the present, Chapter 2 + Other Excerpts, Routledge.

McLaren B, Scheuer O, De Laat M, Hever R, de Groot R, Rosé CP. (2007). Using machine learning techniques to analyze and support mediation of student E-discussions, In: Proceedings of the 2007 Conference on Artifcial Intelligence in Education: Building Technology Rich Learning Contexts That Work, 331–338.

McNamara DS, Graesser AC. (2012). Coh-Metrix: an automated tool for theoretical and applied natural language processing, In: McCarthy PM, Boonthum C eds. Applied Natural Language Processing: Identifcation, Investigation, and Resolution, Hershey, PA: IGI Global.

Milligan, S. (2015). Crowd-Sourced Learning in MOOCs: Learning Analytics meets Measurement Theory, in Proceedings of Learning Analytics and Knowledge 2015.

Molenaar, I. & Chiu, M. (2015). Effects of Sequences of Socially Regulated Learning on Group Performance, in Proceedings of Learning Analytics and Knowledge 2015.

Morrow, R. & Brown, D. (1994). Deconstructing the Conventional Discourse of Methodology: Quantitative versus Qualitative Methods, in Morrow & Brown (Eds.) Critical Theory and Methodology: Contemporary Social Theory Volume 3, Sage Publications.

Mu J, Stegmann K, Mayfeld E, Rosé CP, Fischer F. (2012). The ACODEA framework: developing segmentation and classifcation schemes for fully automatic analysis of online discussions. Int J Comput Support Collab Learn:285–305. 138.

Nehm R, Ha M, Mayfeld E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol 21:183–196.

Nguyen, D., Dogruöz, A. S., Rosé, C. P., de Jong, F. (in press). Computational Sociolinguistics: A Survey, *Computational Linguistics*.

O'Donnell, A. & King, A. (1999). Cognitive Perspectives on Peer Learning, Routledge.

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2009). Contemporary Linguistics: An Introduction, Bedford/St. Martins.

Page EB. (1966). The imminence of grading essays by computer. Phi Delta Kappan 48:238–243.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2(1-2), pp1-135.

Ramesh, D. Goldwasser, B. Huang, H. Daum´e III, and L. Getoor (2013). Modeling learner engagement in moocs using probabilistic soft logic. In Workshop on Data Driven Education, Advances in Neural Information Processing Systems.

Ribeiro, B. (2006). Footing, positioning, voice: Are we talking about the same thing? in Fina, A.,

Schiffrin, D., & Bamberg, M. (Eds). Discourse and Identity, Cambridge University Press

Rosé CP, Tovares, A. (2015). What sociolinguistics and machine learning have to say to one another about interaction analysis. In: Resnick L, Asterhan C, Clarke S, eds. Socializing Intelligence Through Academic Talk and Dialogue. Washington, DC: American Educational Research Association.

Rosé, C. P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F., (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning, submitted to the International Journal of Computer Supported Collaborative Learning 3(3), pp237-271.

Rosé CP, VanLehn K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. Int J AI Educ 15:325–355.

Sekiya, T., Marsuda, Y., Yamaguchi, K. (2015). Curriculum Analysis of CS Departments Based on CS2013 by Simplified, Supervised LDA, in Proceedings of Learning Analytics and Knowledge 2015.

Shermis MD, Burstein J. (2013). Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York, NY: Routledge.

Shermis M, Hammer B. (2012). Contrasting state-of-the-art automated scoring of essays: analysis. In: Annual National Council in Measurement in Education Meeting, 14–16.

Simsek, D., Sandor, A., Buckingham-Shum, S. (2015). Correlations between Automated Rhetorical Analysis and Tutor's Grades on Student Essays, in Proceedings of Learning Analytics and Knowledge 2015.

Skrondal, A. & Rabe-Hesketh, S. (2004). Interdisciplinary Statistics: Generalized Latent Variable Modeling: Multi-Level, Longitudinal, and Structural Equation Models, Chapman & Hall/CRC

Snow, E., Allen, L., Jacovina, M., Perret, C., & McNamera, D. (2015). You've got style: Writing flexibility across time, in Proceedings of Learning Analytics and Knowledge 2015.

Soller, A., & Lesgold, A. (2000). Modeling the Process of Collaborative Learning. In: Proceedings of the International Workshop on New Technologies in Collaborative Learning, Japan: Awaiji–Yumebutai.

Wen M, Yang D, Rosé CP. (2014a). Sentiment analysis in MOOC discussion forums: what does it tell us? In: Proceedings of Educational Data Mining, 2014a.

Wen M, Yang D, Rosé D. (2014b). Linguistic reflections of student engagement in massive openonline courses, In: Proceedings of the International Conference on Weblogs and Social Media.

Witten, I. H., Frank, E., and Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques, third edition, Elsevier: San Francisco