

# Semantic Representation Analysis: A General Framework for Individualized, Domain-Specific and Context-Sensitive Semantic Processing

Xiangen Hu<sup>1,2</sup>, Benjamin D. Nye<sup>1</sup>, Chuang Gao<sup>2</sup>, Xudong Huang<sup>1</sup>,  
Jun Xie<sup>1</sup>, and Keith Shubeck<sup>1</sup>

<sup>1</sup> The University of Memphis, Memphis, TN  
{xhu, bdnye, xhuang3, jxie2, kshubeck}@memphis.edu

<sup>2</sup> Central China Normal University, China  
chuanggaocnucnu@gmail.com

**Abstract.** Language agnostic methods for semantic extraction, encoding, and applications are an increasingly active research area in computational linguistics. This paper introduces an analytic framework for vector-based semantic representation called semantic representation analysis (SRA). The rationale for this framework is considered, as well as some successes and future challenges that must be addressed. A cloud-based implementation of SRA as a domain-specific semantic processing portal has been developed. Applications of SRA in three different areas are discussed: analysis of online text streams, analysis of the impression formation over time, and a virtual learning environment called V-CAEST that is enhanced by a conversation-based intelligent tutoring system. These use-cases show the flexibility of this approach across domains, applications, and languages.

**Keywords:** Semantic analysis, language agnostic, domain vocabulary, intelligent tutoring systems.

## 1 Theoretical Basis: Semantic Representation Analysis

As the internet becomes pervasive worldwide, languages other than English will be increasingly common. Chinese, Spanish, French, German, Korean, and many other languages already have significant footholds in the internet, which are likely to grow, since users prefer sites in their native language. Social networks, which account for a large percentage of web traffic, already tend to be almost entirely in native languages. Semantic analysis can be an extremely useful tool, but the emerging internet presents significant challenges for traditional methods. Text corpora exist across many languages, many domains, and many contexts, such as different time periods or applications. This diversity of text corpora has made language-agnostic methods for semantic extraction, encoding, and applications an active research area in contemporary computational linguistics.

SRA provides a general framework for conceptualizing and applying existing semantic extraction/encoding methods, such as Latent Semantic Analysis (LSA) [1], Hyperspace Analogue to Language (HAL) [2], and bound encoding of the aggregate language environment (BEAGLE) [3]. The two key elements of SRA are the vector-representation of the semantics of language entities (words, idioms, phrases, sentences, paragraphs, documents, etc.) and the numerical relations between language entities (such as similarity, relatedness, or semantic overlap). The basic requirements for SRA are that the representations must be language-agnostic and computationally feasible. Hu, Cai, Graesser, and Ventura [4] outlined SRA based on the following assumptions:

**Hierarchical Representation:** Different levels of a language entity may have their semantics represented differently.

**Algebraic Representation:** The semantics of any level of language entities must be capable of being represented numerically or algebraically.

**Computational Aggregation:** The semantics of a higher-level language entity are computed as a function of semantics for its lower-level language entities. Also, at the lowest level of language entities, a numerical semantic comparison measure must exist between any two items (e.g., words).

These three assumptions are the foundation of a general framework underlying most existing semantic extraction/encoding methods. The hierarchical assumption and the algebraic representative assumption work together to ensure that the language entities can be computed mathematically. This final assumption emphasizes the idea that comparisons occurring at the most basic level should be inputs for higher levels (e.g., the similarity of paragraphs should consider the similarity between their constituent sentences). These assumptions come from considering semantic regularities. First, the five basic language entities are hierarchical: words, phrases, sentences, paragraphs, and documents are each constituted by more basic entities. Second, numerical and algebraic representations for each language entity have been created in recent decades. Semantic extraction/encoding methods (e.g., LSA & BEAGLE) have been used to numerically represent all five levels of language entities, through the creation of semantic spaces. Third, in these semantic spaces, a larger language entity (e.g., document) can be represented by aggregating the semantic relationships of smaller entities (e.g., cosine similarity of words in LSA).

An important concept that is derived from the SRA framework is the Induced Semantic Structure (ISS). ISS focuses on numeric relations between language entities while intentionally de-emphasizing the encoding details (the vector-representation) for the semantic spaces. ISS considers a target word and an ordered list of its top nearest neighbors in a semantic space [4]. At the lowest levels (typically words), one intuitive measure for the semantic similarity can be the unordered overlap between their nearest neighbors. By computing the information from the top nearest neighbors, ISS captures the similarity of the target words, which can be aggregated to calculate the semantic similarity of the higher-level language entities.

The words of J.R Firth [5] best capture the concept of nearest neighbors: “You shall know a word by the company it keeps” (p. 11). This view has been accepted as

an important hypothesis in the field of vector-based semantic analysis: A word's nearest neighbors represent the meaning of the target word. Prior studies have applied nearest neighbors to compare semantics. Andrews, Vigliocco, and Vinson [6] randomly chose words in several spaces and compared their top several nearest neighbors. In this study, different neighbors for the same target word indicated different meanings for the word in the two semantic spaces. The overlap between nearest neighbors has also been used to identify words whose meanings vary across domains [7].

SRA used with ISS is capable of comparing a variety of semantic spaces. These semantic spaces may differ in their encoding methods (such as LSA, HAL), corpora (e.g., Wikipedia, TASA), or parameters (e.g., the number of dimensions in the vector representation). Semantic spaces can be compared to any semantic representations with nearest neighbors, even those not generated algorithmically. A particularly notable semantic representation of this type is the set of free association norms manually collected by humans [8], which is often used as a "gold standard" for semantic meaning.

An evaluation between five Touchstone Applied Science Associates (TASA) spaces [9] was recently conducted and showed that measuring semantic spaces with SRA captured underlying patterns from the corpora effectively [10]. The TASA corpus consists of proper English written text, including textbooks from first grade to the first year of college, with each of the five TASA spaces being additive (e.g., the 6-th grade space includes grades 1-6). The additive property of the spaces imposes natural similarity relationships between spaces. Specifically, two spaces with the largest proportion of document overlap should have the highest semantic similarity. Three distinct space comparison measures based on ISS each successfully captured this similarity pattern [10].

Furthermore, SRA combined with ISS provides an efficient way to create a domain-specific semantic processing portal that is capable of computing semantic relevance to customized domains. Such a capability makes it possible to "decompose" the semantics of any language entity on a list of domains, similar to the way spectrum-analysis does in physics. In the next sections, we will present the Domain-Specific Semantic Processing Portal (DSSPP) web service ([about.dsspp.com](http://about.dsspp.com)) that implements this functionality and a tutoring system based on Sharable Knowledge Objects (SKO; [www.skoonline.org](http://www.skoonline.org)) that consumes the DSSPP web service. Applications of DSSPP are presented, including analysis of online text streams, analysis of the impression formation over time, and a virtual learning environment called V-CAEST that is enhanced by a conversation-based intelligent tutoring system.

## 2 The Domain-Specific Semantic Processing Portal

A proof of concept implementation of SRA has been implemented. This implementation is called the Domain-Specific Semantic Processing Portal (DSSPP). DSSPP is a web-service implemented in a cloud computing platform (Google App Engine and Amazon Elastic Compute Cloud). DSSPP provides web services to 1) compute

nearest neighbors for available semantic spaces, 2) compute semantic relations (e.g., similarity) between any two pieces of English texts within or between two semantic spaces, and can also 3) perform a semantic “spectrum analysis” (e.g., relevance to different domains), and 4) calculate learner’s characteristics curves (LCC) for student statements in a tutoring system. These functionalities can be used for a number of applications, and will be described in the context of the applications where they have been used, to help ground the discussion.

## 2.1 Real-Time Analysis of Topic Evolution in Online Text Streams

SRA has been used to analyze topic evolution in online text streams. This application collected a series of corpora from online text streams such as tweets or online blogs. Each corpus is a slice of the continuous stream of tweets or entries, which will be referred to as the smallest independent corpus (SIC). These SIC are indexed by their occurrence (e.g., time, location). For each SIC, a small semantic space is created and an ISS is extracted for a limited lexicon (i.e., a specific topic of interest). The relationships between the terms for the selected lexicon are analyzed as the function of their indexes (e.g., time) and relevance to a domain.

This technique is useful for studying online social networks, which are influential in contemporary society. Online social networks continuously generate text streams over time, which carry a high volume of information and change quickly. Online text streams have been used to explore public opinion, such as sentiment towards political candidates [11], and customer attitude toward commercial products [12]. Researchers often analyze public opinion in text streams by studying topic evolution [13,14,15, 16]. In this earlier research, a topic is defined as a term or a group of terms and their relations to their referent topics. Therefore, topic evolution is defined as the change in relationships between a topic and its referent topics as a function of internet time [17]. For example, if the topic is education, it semantically relates to teachers, schools, students, knowledge, and other topics which are the referent topics of education. Thus, the topic evolution of education would be the change of semantic relationships between education and its associated topics over a time interval of interest.

Based on the definition of topic evolution in prior work, two issues exist. First, it is not domain specific. Each topic is composed of various referent topics, which may be connected to different domains. Accordingly, the change of semantic relations between a specific topic and its referent topics reflects the change of semantic relations between it and all of its related domains. Kleinberg [18] noted that domain knowledge can be used to interpret the temporal patterns in topic evolution. However, previous studies seldom conduct domain-specific topic evolution [19]. Second, most existing semantic methods do not track topic evolution using online algorithms. Instead, they employ retrospective analysis to consider topic evolution. When new texts arise, a retrospective approach adds the new texts to the old ones to update the parameters of a model in order to generate the new trend of topic. This approach can incur resource-intensive computation that is hard to perform in real-time. To address these issues, a new method based on DSSPP was applied to generate topic evolution as a function of domains in real time [17].

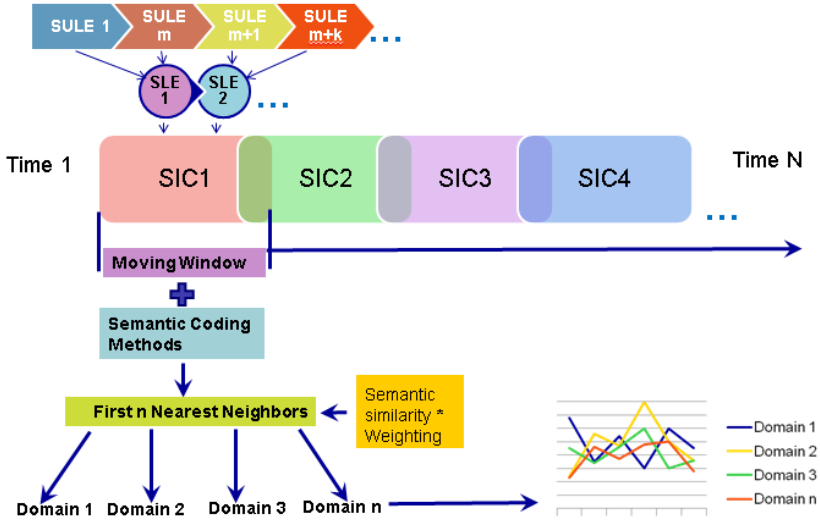


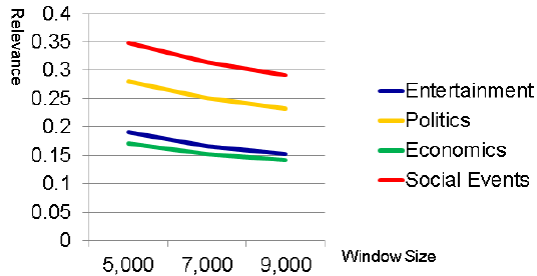
Fig. 1. The process for tracking topic evolution using DSSPP [17]

In this approach, online text streams are decomposed into three levels: 1) the Smallest Unit of Language Entity (SULE), which is usually a word or some special combination of words; 2) the Smallest Language Environment (SLE), which are constituted by SULE (such as a tweet); and 3) the Smallest Independent Corpus (SIC), which is the highest level and composed by SLE. The moving window, which is a time frame sliding on the timeline, generates temporal SICs. Therefore, online text streams are processed as a sequence of time-ordered SICs. In each SIC, semantic analysis is applied to generate nearest neighbors for the topic. According to their semantic similarity to the topic, the top  $N$  neighbors represent the semantic associations for the current SIC. Then, these top  $N$  neighbors are analyzed based on topic-related domains, where each domain contains a selected set of words and their semantic relationships. This method computes the topic's relevance to each domain based on the number of overlapping top  $N$  nearest neighbors between the SIC's semantic space and that domain's semantic space. If desired, the neighbors' importance to the topic can also be considered, by taking the order of the nearest neighbors into account. After calculating the relevance to the domains, the moving window slides to create a new SIC and the topic relevance to each domain is continuously calculated. This produces a time series for the relevance of each SIC to each domain, which is the topic evolution with respect to each domain. Figure 1 displays the process for this method for tracking topic evolution in real time.

This new method was applied to a topic on a serious car accident. The data came from Sina Weibo, which is China's equivalent to Twitter. As such, Chinese semantic spaces were generated and analyzed using DSSPP. In this application, a moving window generated three sizes of SICs in order to test the effect of window size on the method's performance. These sizes were 5000 documents, 7000 documents, and 9000 documents per SIC window. Latent semantic analysis was employed to generate

nearest neighbors for the topic. The top 1000 neighbors were used to analyze the topic’s meaning in each SIC. The topic’s relevance was computed with respect to four domains: politics, social events, economics, and entertainment.

The goal of this application was to test the effectiveness of the new method and examine the influence of SIC size on the method’s performance. Preliminary results using this method indicated that the topic’s relevance to the social domain was significantly higher than other domains. This follows what was anticipated, since the car accident generated significant debate over social issues, but was not particularly tied to economics or entertainment. There was, however, some political debate over the role of the police. As shown in Figure 2, the results were stable across the different SIC window sizes. Future work is planned to compare the topic evolution patterns against human ratings for this corpus, to help validate this method.



**Fig. 2.** Topic relevance to domains across SIC sizes [17]

## 2.2 Semantic Analysis to Track Impression Formation

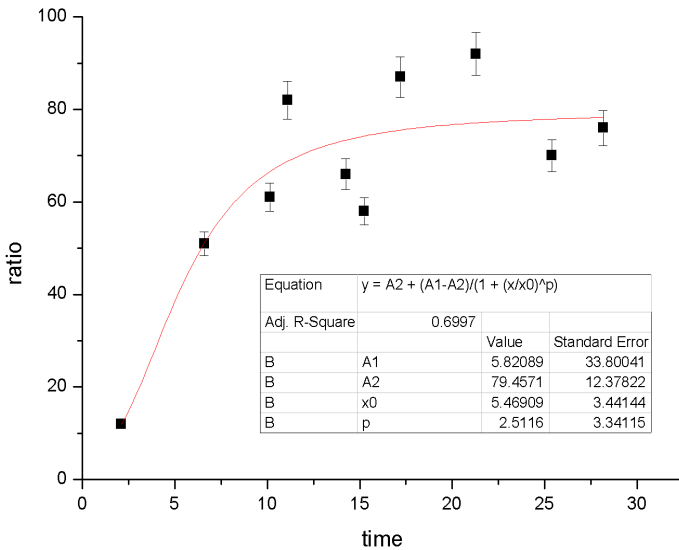
A second application of SRA is the “semantic decomposition” capability. Within SRA, any piece of texts can be semantically “decomposed” based on a customizable set of domains or broken down based on the sentiment expressed (e.g., negativity versus positivity as domains). A case study of impression formation on the internet shows the capabilities of this approach. In social psychology, impression formation is the process that integrates separate pieces of information about a person into a holistic and global impression of that person. In online social networks, news stories about famous men or women are constantly being published. After publishing a story, individuals in certain areas of the world (e.g., China) send off Bulletin Board System (BBS) texts about these stories. The news is connected with the BBS text content. So then, BBS texts carry important semantic information about an individual concept.

To examine this phenomenon, a story was analyzed as an event associated with a sample of BBS texts from different subjects. A three step process collected semantic data on impression formation with respect to an individual news story and its subsequent BBS texts. First, the topic was selected. For this case study, we chose a famous

Chinese internet model. The number of BBS texts for each story ranged from 630 to 25400. This subject was chosen for two reasons:

1. All the news on this model is largely on the same topic. Though there are differences in the level of focus, all stories focus on the same class of events (stimuli);
2. A corpus of stories and data on the model could be collected starting from the beginning of the model’s celebrity status until the present day.

Second, using semantic analysis, keywords for each text were classified into three categories of sentiment: positive, negative, and neutral. As the stories and BBS texts were in Chinese, this analysis was performed using semantic analysis based on Chinese semantic spaces. Third, the ratio of negative keywords was calculated for each BBS text to form a time series. Each news story was considered as a psychological event or stimuli, forming one time point in the time series.



**Fig. 3.** The time series for impression formation

Figure 3 shows the time series of stories on the given celebrity model over time. Each data point is one story. The ratio shown is the average proportion of negative sentiment over the large sample of BBS texts for each story. In this case, sentiment was initially positive, progressed to neutral fairly quickly, after which it became increasingly negative. A logistic regression was fit to this data, which had a moderate fit. The rising behavior shows the changes that occur during the formation of the groups’ impression of the model, which appears to stabilize around fairly negative opinion. Further exploration is looking into the rates for impression formation to stabilize for different types of topics, in order to see if certain topics tend to stabilize more quickly (or not at all).

### 2.3 Learner's Characteristic Curves (LCC) to Drive Tutoring Dialog

Finally, the semantic processing web service provided by DSSPP can be used by Sharable Knowledge Object (SKO) modules, an implementation of AutoTutor [20]. AutoTutor is an intelligent tutoring system (ITS) framework developed at the University of Memphis over the last 17 years. SKO uses two functions of DSSPP. Semantic similarity is computed to evaluate overlap between a student's natural language input and expected ideal answers for questions. Across multiple student inputs, learner's characteristic curves (LCC) are calculated to track the novelty and relevance of student's free recall of learned content in a self-elaboration style interaction. These LCC curves are used to determine the appropriate feedback to present to the student.

An AutoTutor module typically consists of two or more animated agents, where one agent represents a tutor. The tutor agent guides the learner through various domain concepts by applying a conversational framework based on constructivist theories of learning [21] and the behavior of expert human tutors [22]. Empirical evaluations have shown that AutoTutor provides an effective learning environment, with learners using AutoTutor averaging about  $0.8\sigma$  higher learning gains over control conditions such as reading static text [23]. Recently, AutoTutor has moved towards modularity by adopting a Sharable Knowledge Object (SKO) framework. A SKO uses natural language processing and dialog engines hosted on a cloud server. Distributing static content and media on cloud servers allows AutoTutor to be broken down into its key components [24].

AutoTutor Lite takes advantage of this modular, cloud-hosted, SKO framework by providing a web-based tutoring system which uses some, but not all, of the components of AutoTutor. AutoTutor Lite includes the AutoTutor-style interface and conversational framework, incorporating animated agents and natural language [25]. AutoTutor Lite also contains a simplified authoring tool which allows subject matter experts or instructors to create effective learning modules with minimal computer skills. A typical AutoTutor Lite module contains several "slides" dedicated to information delivery and each can contain various media (images, video clips, sound clips, etc). The information delivery section is typically followed by some form of knowledge assessment. AutoTutor Lite allows authors to choose from several assessment types, including fill-in-the-blank, matching, multiple choice, self-reflection and tutoring. The fill-in-the-blank, matching and multiple choice types are best at assessing shallow knowledge and recall, while self-reflection and tutoring assessment types are designed to assess and reinforce qualitative or conceptual knowledge.

AutoTutor Lite assesses student responses in real time by using a lightweight language analyzer based on DSSPP. This analyzer creates a simple micro-model of student knowledge referred to as the Learner's Characteristic Curves (LCC). Student responses are evaluated and compared to an ideal answer (expectation) through the use of semantic analysis provided by a DSSPP. Authors can select between corpora from several domains (science, mathematics, computers and internet, health, etc.), which interpret input using different semantic relationships and domain-specific terms.



The LCC builds curves which describe a student's knowledge on a given topic, based on two metrics: relevance and novelty. Relevance (R) is calculated as the semantic similarity of student input to the ideal answer. Novelty (N) is calculated as the semantic similarity between the student's current input with their history of prior statements. From these, four curves are generated based on the sequence of student responses: Relevant+New ( $N * R$ ), Irrelevant+New ( $N * (1 - R)$ ), Relevant+Old ( $(1 - N) * R$ ), and Irrelevant+Old ( $(1 - N) * (1 - R)$ ). A total coverage score is also calculated, which evaluates the total combined relevance of student statements. When developing AutoTutor Lite modules, authors create specific feedback triggers using these LCC curves. For example, if a student consistently provides irrelevant-old information, the authors can create a trigger that prompts the tutor agent to guide the learner back to the issue at hand, or to suggest a review of the content. These triggers consist of rule-sets contingent on the current and prior values of LCC curves (e.g., if Relevant+New  $< 0.1$ , provide a hint).

An ongoing project called V-CAEST (Virtual Civilian Aeromedical Evacuation Sustainment Training) takes advantage of the SKO framework with AutoTutor Lite and its LCC student model. The central goal of V-CAEST is to improve communication between civilian medical practitioners and military personnel during disaster situations (e.g., Hurricane Katrina). To accomplish this goal, a virtual world has been developed using the Unity 3D game engine. AutoTutor Lite is embedded within this virtual world, and helps guide and tutor users in the game world.

To more accurately evaluate user input in the V-CAEST interface, a domain specific semantic space was developed. Both the medical and military fields involve a great deal of domain-specific vocabulary. Most semantic engines are trained on general corpora, such as TASA [26], because these corpora generalize to common English use-cases. However, a semantic engine trained on a general corpus is unsuitable for V-CAEST, which needs to determine the semantic similarity of responses within a specialized domain. Despite the large size of TASA and other spaces, many esoteric medical and military terms and acronyms are not even included in general corpora. To solve this issue, a guided web crawler iteratively collected domain-specific corpus of articles from a source (e.g., a wiki) around a starting set of domain-specific seed terms provided by subject matter experts.

A screenshot of this interface is shown in Figure 4. In V-CAEST, users are situated in a city block recently struck by a large earthquake. They are required to locate and triage several victims. As a user triages victims, they receive just-in-time feedback from AutoTutor Lite. For example, if a user selects an incorrect triage category, an AutoTutor Lite SKO is triggered and helps explain the mistake that was made. This shows how V-CAEST combines four key technologies into a virtual world: a domain specific processing portal, shareable knowledge objects, web-based intelligent tutoring systems, and a lightweight student model (LCC).



Fig. 4. The V-CAEST 3D world with an AutoTutor Lite SKO tutoring dialog active

### 3 Conclusions and Future Directions

As shown in these examples, the Semantic Representation Analysis and the Domain-Specific Semantic Processing Portal (DSSPP) have a variety of useful applications for research and educational applications. In particular, it is important to note the breadth of the language that this approach can accommodate: the examples described here cover general English (TASA corpora), informal Chinese (Weibo, BBS), and constrained military-medical English terminology. English and Chinese can both be handled using this approach, with the difference being that Chinese requires a parser that segments characters into words (as word boundaries are less clear than English). In future work, we hope to apply DSSPP to additional languages and domains.

Significant work remains for exploring the current directions described in this paper as well. The V-CAEST project is about to begin two phases of evaluation: expert evaluation of the military-medical semantic space using a triad task (i.e., selecting which of two sentences are more similar to an exemplar sentence) and evaluation of learning outcomes for subjects using the V-CAEST environment. These should provide insight into how learners speak and learn in virtual worlds. Further study of online feeds and streams is also continuing, focusing on how a topic's relevance to different domains evolves over time. In particular, relationships between domains may prove an interesting area of study. If language about certain topics goes through certain discrete phases of focus (e.g., it initially focuses on social events, then shifts toward politics), these patterns could be important for anticipating and understanding discourse transitions in online environments. Future work will also focus on validating the computed levels of domain-relevance against human judgments made for samples of texts.

**Acknowledgements.** Our thanks to the Office of Naval Research STEM ITS Grand Challenge, the Institute of Education Sciences, and the US Army, whose grants have supported this fundamental research. All statements are the responsibility of the authors alone.

## References

1. Landauer, T., Dumais, S.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2), 211–240 (1997)
2. Burgess, C.: From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, and Computers* 30(2), 188–198 (1998)
3. Jones, M.N., Mewhort, D.J.: Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* 114(1), 1–37 (2007)
4. Hu, X., Cai, Z., Graesser, A.C., Ventura, M.: Similarity between semantic spaces. In: 27th Conference of the Cognitive Science Society, pp. 995–1000 (2005)
5. Firth, J.R.: *Papers in Linguistics 1934–1951*. Oxford University Press, London (1957)
6. Andrews, M., Vigliocco, G., Vinson, D.: Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116, 463–498 (2009)
7. Itagaki, M., Aue, A., Aikawa, T.: Detecting inter-domain semantic shift using syntactic similarity. In: Calzolari, N., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) 5th International Conference on Language Resources and Evaluation, pp. 2399–2402. European Language Resources Association, Paris (2006)
8. Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36, 402–407 (2004)
9. Zeno, S.M., Ivens, S.H., Millard, R.T., Duvvuri, R.: *The educator's word frequency guide*. Touchstone Applied Science, Breward (1995)
10. Huang, X., Nye, B.D., Xie, J., Wang, J., Liao, Y., Hu, X.: Evaluating a method to compare human and machine-generated semantic spaces. Submitted for Review (forthcoming)
11. O'Connor, B., Balasubramanian, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Hearst, M., Cohen, W., Gosling, S. (eds.) 4th International AAAI Conference on Weblogs and Social Media, pp. 122–129. AAAI Press, Menlo Park (2010)
12. Mishne, G., Glance, N.: Leave a reply: An analysis of weblog comments. In: 3rd Annual Workshop on the Weblogging Ecosystem, Edinburgh, Scotland (2006)
13. Lee, C.H.: Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications* 39(10), 9623–9641 (2012)
14. Alsumait, L., Barbará, D., Domeniconi, C.: On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Giannotti, F., Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., Wu, X. (eds.) 8th IEEE International Conference on Data Mining, pp. 3–12 (2008)
15. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream. In: Elmagarmid, A.K., Agrawal, D. (eds.) 2010 ACM SIGMOD International Conference on Management of Data, pp. 1155–1157 (2010)

16. Chen, J., Yu, J., Shen, Y.: Towards topic trend prediction on a topic evolution model with social connection. In: Zhong, N., Gong, Z., Cheung, Y., Lingras, P., Szczepaniak, P., Suzuki, E. (eds.) *IEEE International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 153–157 (2012)
17. Xie, J., Hu, X.: A method to explore topic evolution in online text streams. Masters Thesis manuscript, University of Memphis (2014)
18. Kleinberg, J.: Temporal dynamics of on-line information streams. *Data Mining Knowledge Discovery* 7(4), 373–397 (2006)
19. Ganesh, M., Reddy, C., Manikandan, N.: TDPA: Trend Detection and Predictive Analytics. *International Journal on Computer Science and Engineering* 3(3), 1033–1039 (2011)
20. Graesser, A.C., Chipman, P., Haynes, B., Olney, A.: AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48(4), 612–618 (2005)
21. Aleven, V., Koedinger, K.: An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science* 26(2), 147–179 (2002)
22. Graesser, A.C., Person, N.K.: Question asking during tutoring. *American Educational Research Journal* 31(1), 104–137 (1994)
23. Graesser, A.C., D’Mello, S., Hu, X., Cai, Z., Olney, A., Morgan, B.: AutoTutor. In: McCarthy, P., Boonthum, C. (eds.) *Applied Natural Language Processing and Content Analysis: Identification, Investigation and Resolution*, pp. 169–187. IGI Global, Hershey (2012)
24. Nye, B.D.: Integrating GIFT and AutoTutor with Sharable Knowledge Objects (SKO). In: Sottolare, R.A., Holden, H.K. (eds.) *Artificial Intelligence in Education (AIED) 2013 Workshop on GIFT*, pp. 54–61. CEUR (2013)
25. Hu, X., Cai, Z., Han, L., Craig, S.D., Wang, T., Graesser, A.C.: AutoTutor Lite. In: Dimitrova, V., Mizoguchi, R., Du Boulay, B., Graesser, A.C. (eds.) *Artificial Intelligence in Education (AIED)*, p. 802. IOS Press, Amsterdam (2009)
26. Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantic analysis. *Discourse Processes* 25, 259–284 (1998)