Evaluating Learning Language Representations

Jussi Karlgren[1][2], Jimmy Callin[1], Kevyn Collins-Thompson[3], Amaru Cuba Gyllensten[1], Ariel Ekgren[1], David Jurgens[4], Anna Korhonen[5], Fredrik Olsson[1], Magnus Sahlgren[1], and Hinrich Schütze[6]

[1] Gavagai, Stockholm
[2] Kungl Tekniska Högskolan, Stockholm
[3] University of Michigan, Ann Arbor
[4] McGill University, Montréal
[5] University of Cambridge
[6] Ludwig-Maximilians-Universität, München

Abstract. Machine learning offers significant benefits for systems that process and understand natural language: a) lower maintenance and upkeep costs than when using manually-constructed resources, b) easier portability to new domains, tasks, or languages, and c) robust and timely adaptation to situation-specific settings. However, the behaviour of an adaptive system is less predictable than when using an edited, stable resource, which makes quality control a continuous issue. This paper proposes an evaluation benchmark for measuring the quality, coverage, and stability of a natural language system as it learns word meaning. Inspired by existing tests for human vocabulary learning, we outline measures for the quality of semantic word representations, such as when learning word embeddings or other distributed representations. These measures highlight differences between the types of underlying learning processes as systems ingest progressively more data.

Keywords: language representations, semantic spaces, word embeddings, machine learning, evaluation

1 Introduction and Motivation

For language technologies that need to represent and understand the meaning of text, machine learning provides a crucial tool for supporting new terminology or semantic interpretations. Learning allows systems to adapt without the need to manually curate knowledge bases, thereby lowering maintenance costs, and to quickly retrain for new domains. Indeed, new genres and communication channels make the requirements for an adaptable system all the more greater [6]. However, learning comes at a risk: The behaviour of an adaptive resource is less predictable than that of an edited stable resource, and quality control thus becomes a continuous issue, rather than something which is done when a knowledge resource is deployed.

Currently, many existing natural language and information retrieval systems that employ learned semantic representations are evaluated after learning completes. While these tests are effective at measuring performance on a specific task, application, or domain, they capture only the outcome of learning while

probe item	*correct answer*	*confounder items*
haphazardly	randomly	linearly densely dangerously

Table 1. An example TOEFL synonym-selection item

providing little insight into the learning process itself. Thus, a tailored solution may be able to achieve high scores on an outcome-oriented test without measuring the advantage of introducing learning. While this outcome-based evaluation does reflect the motivation for performing a task well, it does not measure the specific aspects which might make a particular learning technique worth re-using.

We propose a new evaluation benchmark aimed at measuring the process of learning, which enables capturing phenomena such as adaptability to new data, sensitivity to the order of example data, and the rate of learning. As a case study, we outline three tests for evaluating the learning process when creating semantic word representation, e.g., the word embeddings produced by word2vec [8]. Such representation are widely used in language technology and must capture a wide variety of meanings [9]. Our evaluation builds upon existing outcome-oriented metrics to illuminate the role and impact of learning.

2 Testing Outcome Versus Process

Techniques for learning word meaning typically process many examples of a word's usage to arrive at a representation of its meaning. While many representations are opaque to direct interpretation (e.g., dimensionally-reduced vectors), the quality of these representations may nonetheless be evaluated by comparing the representations themselves, where words with similar meanings are expected to have similar representations. Thus, the most common tests involve testing various aspects of synonymy between terms, with a frequent benchmark being the TOEFL test [7] which consists of a set of target words and a multiple-choice set of options for each from which the best synonym should be chosen, as shown in Table 1. The TOEFL test is typically applied by presenting a respondent with a probe item and some candidates from which the correct item is chosen. This means that the system may be able to answer correctly without ever having established any relationship between the probe item and the correct answer and that the test does not measure the quality of the semantic neighbourhood or semantic field the system has learnt.

Later and more fine-grained tests have included multiple relationship types. For example, the BLESS test divides up the general relation of semantic association into specific relationships such as synonymy, hyponymy, or meronymy between the probe word and the test items. This allows for more detailed analyses of semantic similarity. The authors explicitly state that their intention was to enable testing on specific and intrinsic characteristics of the testable representations under consideration [1, 5].

While these types of outcome-based tests offer valuable contributions for differentiating the qualities of semantic representations, we propose a different but complementary objective that assesses qualities of the *learning process*, not only the final learning outcome. Such *process-based testing* would evaluate *how* various models progress toward learning representations with the qualities that they are intended to capture. At their simplest, process-based tests could be performed by applying an outcome test at intervals throughout a learning process to track the progress of learning the set of probe terms; more sophisticated designs may incorporate insights from developmental psychology or learning theory when creating test items.

Process-based testing would detect potential differences between representation-learning approaches. For example, one model might be designed to learn representations that capture all the diversity of a word's meanings, whereas another may be designed to converge to a representation for the most-frequently seen meaning as quickly as possible; whereas the final representations of both models may produce similar results with outcome-based testing, process-based testing may highlight cases where one model would be preferred over the other, e.g., quicker convergence. Furthermore, given the recent interest in computationally-intensive models such as word2vec [8], an evaluation benchmark which assesses the learning process itself will be of practical utility for understanding the learning rate and representation robustness as more training data is seen.

3 Existing tests for human language learning

In designing a process-based evaluation for automated language learning, we can draw on recent related progress in cognitive psychology that has developed methods for evaluating the human language learning process - and more specifically, on tracking how people acquire new vocabulary. The human learning process for vocabulary is incremental: it involves knowledge of individual words that is often passive, unstable, and partial [3]. Human vocabulary competence has been tested in a variety of settings that include reading comprehension, synonym judgment, synonym generation, gap filling and cloze exercises, acceptability assessment, paired analogies, and translation or paraphrasing. However, traditionally there had been little work on sensitive assessment measures that could detect the partial and incremental aspects of the word learning process. Thus, a key inspiration for our computational work is a recently-developed line of research in contextual word learning that tracks incremental changes and improvements in word knowledge as people are exposed to words in different contexts over time [4]. The resulting assessment methods include a form of lexical learning test that controls for numerous characteristics of the sample probe terms and contexts given, most notably the semantic constraints imposed by the context. Probe items are sample sentences of infrequent words, which are presented to human subjects who are then asked to self-assess knowledge of them, verify synonyms or to generate synonym items. Example words are given in Table 2 with contexts of varying semantic constraint levels. The type and level of these constraints can be computed and calibrated via crowdsourcing of cloze assessments or other semantic judgment tasks.

4 Requirements for a Learning-Focused Evaluation

As an initial case study in how to design process-based tests, we examine the evaluation requirements for the task of learning word-based semantic represen-

Constraint	Target	Test item
High	*cavort*	In winter the dogs frolic and ..... in the snow.
Medium		The monkeys hooted as they ..... in the branches.
Low		Ida and Peggy meet after work to ..... outside.
High	*recluse*	Joanne likes being alone and doesn't trust people because she's a ..... .
Medium		Mandy has twenty cats and no family, a typical ..... .
Low		We weren't able to tell if the man was a(n) ..... or not.

Table 2. Test items shown to human subjects for word substitution under varying semantic constraints [4].

tations. Here, test items consist of comparisons between vocabulary items and measuring the appropriateness of a particular word usage. In addition to the standard requirements when designing lexical tests, such as test items being balanced for word frequency, part of speech, polysemy, and distributional qualities, we propose four desiderata for the items comprising the test set.

1. A test should be robust across the domains and datasets used during learning and not require a specific dataset to be used for training; ideally, such test items should be recruited from the core vocabulary of the language.
2. A test should be sensitive to the task of learning a new meaning for an item it already knows, as well as to learning how a particular item's meaning has adapted over time. This requires the test to be able to show that a representation can handle seeing usages of a known item in a new domain, upholding the distinction between when an item has acquired a new sense versus when it has not changed.
3. Test items should not be biased towards learning a specific kind of representation in order to compare systems with complementary goals, such as rapid or one-shot learning, learning multiple representations for a word's meanings, or learning representations that encode multiple relationships.
4. The intrinsic properties of the test items should be quantified. For example, recording the difficulty of test items (e.g., as measured by human performance) enables assessing whether systems correctly answer easy items first during learning or whether mistakes occur randomly across the dataset.

Following, we propose three tests and then outline the general testing procedure.

Paradigmatic Usage Test. The first test consists of evaluating context-independent paradigmatic usage, similar to TOEFL [7] and BLESS [1]. Test items are constructed by giving a probe word for which the system must identify which word has the desired semantic relation from a list. Underlying this test is the notion that as a model learns the representation for a word, those words that are semanticly similar would begin to have similar representations, i.e., appear in the word's semantic field. To control for the effects of polysemy and relationship interpretation, we propose selecting probes from relatively closed semantic fields: colour names, names of months, names of countries, professional roles, categories of animals. These probe items' linguistic properties, e.g., relative frequency and polyseymy, can then be measured to create a representative test set where confounder items have similar properties.

health benefits of *coconut* oil include hair care, skin care, and proper digestion and metabolism
the *coconut* tree is a member of the palm family
in the rendang beef stew from sumatra, chunks of beef are cooked in *coconut* milk along with other spices
thanks to a promotion from the airline you can now book a *coconut* to frankfurt for 100 off
he looks dapper in a *coconut* as he arrives for the emporio armani show during milan fashion week
it's on the 28th floor of the *coconut* and it's got all the charms of a corporate headquarters

Table 3. Example test items for the word *coconut* from the Plausible Utterance Test drawn from human-generated text (top) and term substitution (bottom).

Plausible Utterance Test. The second test embeds the same lexical items of the first test in contexts, some of which have been found in naturally occurring text and some of which have been generated through replacing some unrelated word with a probe word, in effect generating implausible contexts of use. The target task is to rank the samples in order of plausibility, ideally ranking confounder items as least plausible. Table 3 shows an example of this question type.

Representational Stability and Agility Test. The third test measures the ability of a model to update its meaning representation when observing new data in two condiditions. In the Stability condition, a model is tested on its ability to maitain a self-consistent representation of a word when observing new contexts for that word that have the same meaning but differ in their contextual features, e.g., examples from a new domain; here the representation should not change drastically, as the underlying meaning has not changed. However, in the second Agility condition, a model is tested on how quickly it can adapt a word's representation when the nex contexts contain a new meaning not seen in training contexts, e.g., a novel sense appears. One possibility for creating these test items would be extending tests on identifying novel word senses [2] with confounding words whose meaning does not change but whose surrounding context does.

Test Procedure and Reporting. All three evaluations follows a similar testing procedure. For all tests, a target system is provided with examples of a targeted word, drawn from a corpus according to specific, desired properties (e.g., corpus domain, number of example instances). An evaluation may have the system learn different representations from multiple corpora in order to control for the effect of the corpus itself. During the learning process, the model is tested at desired testing intervals, e.g., after seeing k examples of the a probe item in training.

For the Paradigmatic Usage test, the model is queried for the semantic field of each probe term and given a target set of k related words, a system is scored according to the fraction of items in the semantic field are in the set. For the Plausibility test, the system ranks contexts for the probe word in order of plausibility; scoring calculates how many of the naturally-occurring contexts are ranked higher than the artificially-generated ones. The Stability and Agility tests are measured according to changes in the semantic field of probe words between testing intervals; Stability measures the degree of similarity in the field, whereas Agility measures the percentage of words associated with the probe's new meaning now in the semantic field. Each test's performance is measured with respect to the testing interval of the probe item and reported as learning curves.

5 Conclusion

We advocate the creation of a shared benchmark for lexical learning which evaluates the process of achieving a learning outcome rather than the outcome itself. The proposed benchmark builds upon existing outcome-based tests by controlling for the conditions in which learning occurs, which allows for extending the benchmark to new semantic objectives (e.g., representing antonymy) or to new domains by incorporating additional datasets under the same conditions.

Acknowledgements

This paper reports results from a workshop on Evaluating Learning Language Models which in part was supported by the European Science Foundation, through its ELIAS project for Evaluating Information Access Systems. Parts of this research (Collins-Thompson) were supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140647 to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

1. Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. In *Proceedings of the 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. ACL, 2011.
2. Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. Novel word-sense identification. In *Proceedings of COLING*, pages 1624–1635, 2014.
3. Gwen A. Frishkoff, Kevyn Collins-Thompson, Charles A. Perfetti, and Jamie Callan. Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavior Research Methods*, 40(4):907–925, 2008.
4. Gwen A. Frishkoff, Charles A. Perfetti, and Kevyn Collins-Thompson. Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, 15(1):71–91, January 2011.
5. Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*, 2014.
6. Jussi Karlgren, editor. *Proceedings of the EACL workshop on New Text: Wikis and blogs and other dynamic text sources*. EACL, 2006.
7. Thomas Landauer and Susan Dumais. 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.
8. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, 2013.
9. Peter D. Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.