EUROPEAN
SCIENCE
FOUNDATION

Research Networking Programmes

Science Meeting – Scientific Report

The scientific report (WORD or PDF file - maximum of seven A4 pages) should be submitted online within two months of the event. It will be published on the ESF website.

***Proposal Title*:**
Evaluating Learning Language Models

***Application Reference N°*:**
5590

1) Summary (up to one page)

Learning Language Models are a challenge for systematic evaluation. Information access systems and linguistic resources have traditionally been evaluated and tested by comparison with some gold standard test set of humanly assessed materials. This workshop will explore how learning, and, thus, changing resource can be evaluated systematically and quantitatively to best elucidate their strengths and weaknesses.

challenge questions:

- What evaluation frameworks are currently in use? What are their strengths and weaknesses? What do they measure?
- What aspects of Learning Language Models do we believe are the most important and persuasive in light of future requirements for application and future directions of research? Can those aspects be measured and compared across models and data sets?
- What tasks are the most attractive ones for proving the worth and necessity of Learning Language Models? Where are the dynamic and emerging qualities of such models most useful?

2) Description of the scientific content of and discussions at the event (up to four pages)

Learning Language Models are an increasingly active research direction which in the last few years has proven useful in practical implementation for information access tasks. One family of such models are the Distributional Semantic Models, which have become the state-of-the-art model of choice in computational semantics. The approach includes models such as Latent Semantic Analysis (Landauer and Dumais), Random Indexing (Sahlgren & Karlgren & Kanerva), Topic Modelling (Hofmann, Blei et al) and Deep Learning (Hinton, Bengio, Mikolov). These models are all built on cooccurrence statistics, which is collected, aggregated, and represented variously.

The models are claimed by their originators to acquire and represent anything from purely semantic information, to some opaque mixture of semantic and grammatical information, to associative information, to topical information.

They are claimed to be useful for tasks ranging from such that are internal to the field such as lexicon generation to information-oriented such as language learning, authorship analysis, search and filtering, and various information access activities where more rigid models typically fail.

Most of these claims are difficult to validate or disprove.

Cranfield-style evaluation has during the past few decades been an instrumental, effective, and useful research vehicle for developing new algorithms and representations for practical text analysis and especially its information access applications.

Claims of the first kind are ill-defined, since there is little agreement e.g. as to how “semantic” contrasts with “topical”. Claims of the second kind are difficult to organise into a Cranfield-style framework, since the attractive qualities of a learning model are near impossible to capture in a gold standard batch of test data.

Some example evaluations are being used in the field. The most commonly used evaluation methodology for understanding the intrinsic qualities of Learning Language Models is to compare their performance with various kinds of vocabulary tests, such as synonym tests (e.g. the synonym part of the TOEFL and ESL), association norms (e.g. the University of South Florida Free Association Norms), or specially designed tests for Learning Language Models (e.g. the BLESS test). Also some tasks, such as the Multi-document Summarization or the Knowledge Base Acceleration tasks of TREC or the Topic Detection and Tracking tasks of the Document Understanding Conference are such that a learning system should perform well in them. It can however be argued that several of these tests only capture some aspects of learning and some aspects of meaning. Vocabulary tests evaluate a resource, Cloze tests evaluate the contextual model, BLESS tests and related tests evaluate relations between referents. The task-based tests are often optimised using a static knowledge resource.

To evaluate the usefulness of Learning Language Models some task-oriented test sets of naturally occurring language have been made available, e.g. through SEMEVAL and RepLab. Best results from these evaluation campaigns are most often given by standard machine learning approaches, not specifically designed for language data.

This workshop discussed and examined existing evaluation schemes and how to use them for evaluating the aspects of learning models that seem most attractive and most useful for future research and for future application tasks.

Why is learning useful?

The reason we like learning is that it gives us systems which

a) have lower maintenance and upkeep cost of expensive resources (compared to having human editors)
b) transfer and port easily to different domains, tasks, or languages
c) adapt nicely to personal or situation-specific settings without manual instruction
d) are robust with respect to various data or domains
e) are timely and accept and adapt to change.

We believe learning systems will prove their worth in any task where today human effort is expended to maintain and update a linguistic resource. Typical such tasks are media monitoring, news tracking, or risk assessments of various types.

But any learning system incurs a risk. The behaviour of an adaptive resource is less predictable than that of an edited stable resource. Quality control becomes a continuous issue, rather than something which is done at deploy time of a resource. The challenge is that most existing language technology and information retrieval evaluations are disparate and adhoc, geared to one specific domain or application, and evaluate outcome or the quality of a resulting resource, rather than the learning process. This, of course, is in itself not a bad thing, if the evaluation is a good fit to some task of interest. However, a test on the outcome fails to measure the specific aspects which make a learning system worth using. A well-edited and tailored solution always will be able to achieve high or even perfect scores on any test, with no advantage of introducing learning.

The discussion at the workshop centered on questions such as: what aspects of learning language should be tested - the vocabulary itself, the competence of the resulting system, its competence in associating terms to appropriate contexts, its competence in abstracting from terms to concepts, and its competence in abstracting relations from the contextual or associative closeness of terms or concepts learned.

The consensus was that an appropriate test should measure and assess the quality of the learning process, but there was clear disagreement on whether that process should be assumed to be guided by hypothesised parameters and principles of human information processing or if the process should be black-boxed.

Also, an invitation to discuss the measurement of pragmatics and informational qualities of language use was not met with unqualified approval among the participants, since the challenges associated with such approaches were deemed theoretically too complex for practical experimentation at this date.

3) Assessment of the results and impact of the event on the future directions of the field (up to two pages)

Two concrete outcomes were decided on

First, to collate and make accessible a test suite of established and known useful tests of various types, with the intention of aiding comparison across systems and data sets. This suite should be presented at a workshop for a wider audience. It was decided to investigate the possibility of making this happen at some computational linguistics conference in the near future. For this purpose a shared document will be defined to collect ideas, pointers, links, and other related thoughts.

Second, to implement a test based on existing tests developed for assessing human vocabulary learning. We will define such a test and run it on our respective materials, hoping to see others do the same. The general idea is to expose a system to a sequence of sentences containing some probe word, and after each sample occurrence we will test the system for synonyms for the unknown word. If this results in an exact or near hit, with metrics and grading to be determined separately, we may record how many samples are necessary. This provides a measure of the learning process. Recording which samples provide best evidence or fastest learning will give us a measure of the semantic resolution of that context. When this test runs smoothly, it will be published jointly and the necessary resources will be made available publicly, not least through the test suite discussed above.

Annex 4a: Programme of the meeting

Friday, October 10

- □ Welcome meet-up for those arriving from abroad
- □ Informal dinner for those without jet-lag

Saturday, October 11

- □ tea, coffee, sandwiches
- □ Magnus: Welcome to Gavagai!
- □ Jussi: About the Workshop
- □ Round table: background of participants
- □ Round table: everyone's favourite feature, task, idea which involves learning languag emodels
- □ Magnus: survey of current evaluation approaches
- □ All: ideas for new and better evaluation approaches
- □ lunch, booked in restaurant nearby
- □ All: how to make above ideas for new and better evaluation approaches operational
- □ meet up on Slussen at 1845 for harbour ferry
- □ dinner, booked in restaurant on Djurgården

Sunday, October 12

- □ tea, coffee, sandwiches
- □ Jussi: summing up yesterday's findings
- □ Magnus: proposals for actions
- □ Round table: practicalities surrounding action points
- □ Lunch for those in less hurry
- □ Departure to airport

Annex 4b: Full list of speakers and participants

Organisers:

Jussi Karlgren, Gavagai and KTH, Stockholm
Magnus Sahlgren, Gavagai, Stockholm

Invited speakers:

Kevyn Collins-Thompson, U Michigan
David Jürgens, U Montréal
Anna Korhonen, U Cambridge
Hinrich Schütze, U München

Partipants at large:

Amaru Cuba Gyllensten, KTH and Gavagai, Stockholm
Ariel Ekgren, KTH, Stockholm
Jimmy Callin, U Uppsala and Gavagai, Uppsala
Fredrik Olsson, Gavagai, Stockholm