The CogALex-IV Shared Task on the Lexical Access Problem

Reinhard Rapp
Aix-Marseille Université
13288 Marseille
France
reinhardrapp@gmx.de

Michael Zock
Aix-Marseille Université
13288 Marseille
France
michael.zock@lif.univ-mrs.fr

Abstract

The shared task of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex-IV) was devoted to a subtask of the lexical access problem, namely multi-stimulus association. In this task, participants were supposed to determine automatically an expected response based on a number of received stimulus words. We describe here the task definition, the theoretical background, the training and test data sets, and the evaluation procedure used for ranking the participating systems. We also summarize the approaches used and present the results of the evaluation. In conclusion, the outcome of the competition are a number of systems which provide very good solutions to the problem.

1 Introduction

In the framework of CogALex-IV (co-located with COLING 2014 in Dublin) we invited colleagues to participate in a shared task devoted to the lexical access problem in language production. Our aim was to make a quantitative comparison between different systems based on a shared set of data and using the same evaluation metric.

The lexical access problem is very relevant for this workshop series as the quality of a dictionary depends not only on its coverage, but also on the accessibility of the information. Put differently, a crucial point of dictionary development is word access by the language producer, an often neglected aspect. Access strategies vary with the task (text understanding versus text production) and the knowledge available at the very moment of consultation (words, concepts, speech sounds). Unlike readers who look for meanings, writers start from them, searching for the corresponding words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related words) and via diverse access routes. Navigation takes place in a huge conceptual lexical space, and the results are displayable in a multitude of forms (e.g. as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

Given a great number of possibilities of approaching the lexical access problem, we felt that for a competition it was necessary to narrow down the choices in order to be able to come up with a clear task definition. Therefore the CogALex shared task focused on a crucial subtask, namely *multi-stimulus association*. What we mean by this is the following. Suppose we were looking for a word matching the following description: *tasty nut with hard shell originally from Australia*, but could not retrieve the corresponding and intended form *macadamia*. This is the well known tip-of-the-tongue problem where an author knows the word but fails to access its form, even though he is able to retrieve certain features of it (meaning, sound, syllables, ...). People being in the tip-of-the-tongue state always remember something concerning the elusive word (Brown & Mc Neill, 1966). This being so, it would be nice to have a system accepting this kind of information as input, and which then proposes a number of can-

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 1–14,
Dublin, Ireland, August 23, 2014.

didates which ideally should contain the target word. Given the above example, we might enter *tasty*, *nut*, *hard*, *shell*, and *Australia*, and the system would be supposed to come up with one or several associated words such as *macadamia*, *walnut*, *cashew*, or *coconut*.

This paper is meant to provide an overview on the shared task and on its results. It is organized as follows: Section 2 gives some background concerning the theory of word finding. Section 3 describes the task definition and Section 4 the training and the test data sets and the evaluation procedure. Section 5 lists the participating systems, tries to characterize the different approaches, and presents the results. For all systems but one, further details are given in the separate papers (in these proceedings) as provided by the members of the participating groups. Section 6 summarizes the conclusions.

2 The problem of word finding

One could imagine many kinds of shared tasks within the framework of the CogALex workshop. Yet, we have focused here on a very specific problem, namely word finding. To this end we have defined a task demanding participants to come up with a system able to compute reversed word associations. While in the standard association experiment people are asked to provide the associations coming to their mind given some stimulus (prime), we have reversed this situation. Given a set of associations, the system was supposed to predict its trigger. More concretely speaking, participants were given 2000 sets of words, each set containing five words. The task was to determine automatically the sixth element, i.e. the prime (or stimulus), evoking the five words. One could object that this task does not really address the word access problem or its solution, but this is not quite so as we will try to show.

In particular, it seems quite reasonable to claim that an association network with bi-directional links (see Rapp, 2014) is a suitable resource to support word 'finding'. Since words are connected via bi-directional links either of the connected items can be the source or the target during the search (or during navigation).

Although systems designed for the shared task can have many applications (see Section 6), a prototypical one is the tip-of-the-tongue problem, which is a special case (yet a quite frequent one) of word access. So let us briefly describe this problem and the steps needed to overcome it.

One of the most vexing problems in speaking or writing is that one knows a given word, yet fails to access it when needed. Suppose, you were looking for a word expressing the following ideas: *superior dark coffee made of beans from Arabia*, but could not retrieve the intended word *mocha*. What will you do in a case like this? You know the *meaning*, you know how or when to *use* the corresponding word, and in principle you even seem to know its spoken or written *form*, since you have used it some time ago (for more details, see Zock et al., 2010). Yet for some unknown reason you simply cannot access it at the very moment of writing or speaking. The just described situation is called *anomia* or *dysnomia*, which in less technical terms means that a person has a word finding problem. This case is often assimilated with the tip-of-the-tongue phenomenon, which technically speaking is not quite correct, but this shall not concern us here.[1]

To resolve the problem, one can think of many strategies. For example, one can ask somebody, by providing him some hints (cues) hoping that the person can guess the elusive word. Such hints could take various forms like a description (definition or circumlocution), an association or the role played by the target word, say, *instrument used for eating Chinese food* when searching for *chopsticks*. Of course, one can also search in an external resource (dictionary). Unfortunately, most dictionaries are primarily designed for the language recipient and not particularly well suited to assist the language producer. And even if there are quite a number of promising proposals,[2] a lot more could be done these days with the help of corpora, computers, and language technology.

[1] The tip-of-the-tongue phenomenon (http://en.wikipedia.org/wiki/Tip_of_the_tongue) is a weak form of an anomic aphasia (http://en.wikipedia.org/wiki/Anomic_aphasia). Yet, unlike the latter, it is only momentary. It is characterized by the fact that the person (speaker/writer) has only partial access to the word s/he is looking for. The typically lacking parts are phonological (syllables, phonemes). Since all information except this last one seems to be available, and since this is the one preceding articulation, we say: the word is stuck on the *tip of the tongue*.

[2] Think of *Roget's Thesaurus* (Roget, 1852), *WordNet* (Fellbaum, 1998; Miller et al., 1990), Longman's *Language Activator* (Summers, 1993), the *Oxford Reverse Dictionary* (Edmonds, 1999) or *OneLook* which combines a dictionary, WordNet, and an encyclopedia, Wikipedia (http://onelook.com/reverse-dictionary.shtml).

This being said, to build a dictionary for the language producer, certain provisions must be made, and it is easy to understand why. When searching a word form (target), the dictionary user will certainly not search in the entire resource. He will rather navigate in a substantially smaller subset (Zock, 2014; Zock & Cristea, 2014). The question is, how to build this reduced space and how to support then navigation. We will deal here mainly with this first step of search space reduction as it is crucial and this is where associations come into play (Deese, 1965; Cramer, 1968).

The experiments concerning the tip-of-the-tongue problem have systematically shown (Aitchison, 2003; Brown, 1991; Brown & McNeill, 1996) that users being in this state always know 'something' concerning the target word: fragments of the meaning, origin, number of syllables, etc. This being so, any of this could be used to guide the search.

Suppose we focused only on the semantic aspects. In such a case it is reasonable to assume that the target form can be found on the basis of its defining elements (bag of the words contained in the definition). While not being perfect, this works quite well (Dutoit & Nugues, 2002; El-Kahlout & Oflazer, 2004; Mandala et al., 1999; Michiels, 1982). Actually, even Google – although not designed for this – is able to recover in many cases the elusive word. Just try the following example, *spring, typically found in Iceland or in the Yellowstone National Park, discharging hot water and steam*, and chances are that you will find the target word *geyser*. Although not perfect, this is nevertheless quite useful. However, this represents only one kind of cognitive state (knowledge of the definition), and this is certainly neither the only one nor the most frequent one. Indeed, there are many situations where it is hard to come up with a precise definition, and in this case other types of information are used to initiate search, for example, co-occurrences, associations, etc. Hence, if our target is *mocha* it may be accessible not only via its definitional terms (*coffee*, *beverage*, ...) but also via any of its associates: *black*, *hot*, *drink*, *Java*, etc. This is the point where associations come to the centre stage.

Some of the related recently published work has been cited in Rapp (2014), and some other is mentioned by the authors participating in the shared task. Therefore, let us focus here primarily on some of the earlier and nowadays often overlooked related work.

Associative networks have been very popular in Artificial Intelligence at the end of the nineteen-seventies (Findler, 1979). They were proposed to be used for many tasks such as word sense disambiguation, finding brand names, reading between the lines, subliminal communication, brainstorming, and supporting word finding. That is, the tip-of-the-tongue problem is but one of the many possible applications.

The study of associative networks was motivated by the goal to understand the organization of the human memory and the mental lexicon. This led to the building of lexical graphs like WordNet (Fellbaum, 1998), the study of the tip-of-the-tongue problem (Brown & Mc Neill, 1966), error analysis (Fromkin, 1980, 1973) and priming experiments. Priming is said to take place if exposure to one stimulus increases significantly the response to another. Meyer and Schvaneveldt (1971) showed in their seminal experiments that people were faster in deciding that a string of letters is a word when it was followed by an associatively or semantically related word. For example, *nurse* is recognized more quickly following *doctor* than following *bread.* These findings supported also the idea of activation spreading as a method of access or search (Collins & Loftus, 1975).

Associative networks can be considered as a special type of semantic network which were introduced by Richens (1956) and by Ceccato (1956) for quite a different purpose. They were meant to serve as an interlingua for machine translation. These knowledge representation structures were then further developed in the sixties by Simmons (1963) and Quillian (1963, 1966, 1967, 1968, 1969). They finally became famous due to the work done by Quillian and two psychologistst (Collins & Quillian, 1969 & 1970 and Collins & Loftus, 1975). Note that semantic networks can represent language at various levels of granularity: word, sentence (Sowa, 1984) or discourse (Mann & Thomson, 1988). Also, and very relevant for us here is the fact that at the word level, they can represent its semantics, i.e. meaning (Nogier & Zock, 1992), or its place withing the global structure of the mental lexicon (Miller, 1995; Aitchison, 2003; Bonin, 2004). In this latter case words are connected by associations rather than by deep-case roles, and the resulting graphs show word neighborhood (Schvaneveldt, 1989). The fact that the mental lexicon exhibits 'small world' characteristics (http://en.wikipedia.org/wiki/Small-world_network) has been shown by Vitevitch (2008) and by Sporns and colleagues (2004).

For the construction of associative networks knowledge about associations is required. Such knowledge can be obtained in two different ways. One is to ask people what a given term (say *cat*) evokes in

their mind (say *dog*, *mouse*, etc.). Another option is to look at word co-occurrences in corpora, and to derive the associations from them (which, strictly speaking, pre-supposes that the human brain is also doing this). For the purpose of having a gold standard for the shared task, by using the EAT, we have opted for the first possibility. In contrast, most systems constructed by the shared task participants rely on the second.

3 Task definition

The participants received lists of five given words (primes) such as *circus*, *funny*, *nose*, *fool*, and *Coco* and were supposed to compute the word most closely associated to all of them. In this case, the word *clown* would be the expected response. Table 1 shows some more examples.

Given Words	Target Word
gin, drink, scotch, bottle, soda	whisky
wheel, driver, bus, drive, lorry	car
neck, animal, zoo, long, tall	giraffe
holiday, work, sun, summer, abroad	vacation
home, garden, door, boat, chimney	house
blue, cloud, stars, night, high	sky

Table 1. Lists of given words together with their targets.

We provided a training set of 2000 sets of five input words (multiword stimuli), together with the expected target words (associative responses). The way how the datasets were produced will be described in the next section. The participants had about five weeks to train their systems on this data. After the training phase, we released a test set containing another 2000 sets of five input words, but without providing the expected target words.

The participants were given five days to run their systems on the test data,[3] with the goal of predicting the target words. For each system, we compared the results to the expected target words and computed an accuracy based on the number of exact string matches (but without taking capitalization into account). The participants were invited to submit a paper describing their approach and their results.

For the participating systems, we distinguished two categories:

1) *Unrestricted systems*. They could use any kind of data to compute their results.

2) *Restricted systems based on ukWaC*: These systems were only allowed to draw on the freely available ukWaC corpus (Ferraresi et al., 2008)[4] in order to extract information on word associations. The ukWaC corpus comprises about 2 billion words of web texts and provides also lemma and part-of-speech information.

Participants could compete in either category or in both. They were encouraged to further improve on their results outside of the competition after the deadline, and to describe these advances in their papers (in these proceedings).

4 Training and test data sets and evaluation procedure

The training and the test data sets were both derived from the *Edinburgh Associative Thesaurus* (EAT; Kiss et al., 1973). The EAT lists for each of 8400 stimulus words up to 100 associative responses as obtained from test persons who were asked to produce the word coming spontaneously to their mind.

As the EAT uses uppercase characters only, and as this might not suit everybody's needs, we decided to modify its capitalization. For this purpose, for each word occurring in the EAT, we looked up which form of capitalization showed the highest occurrence frequency in the *British National Corpus* (Burnard & Aston, 1998). By this form we replaced the respective word. E.g. *DOOR* was replaced by

[3] The exact dates were: training data release: March 27, 2014; test data release: May 5, 2014; final results due: May 9, 2014.

[4] http://wacky.sslmit.unibo.it/doku.php?id=corpora.

door, and *GOD* was replaced by *God*. This way we hoped to come close to what might have been produced during compilation of the EAT if case distinctions had been taken into account.[5] Since this method is not perfect, e.g. words often occurring in sentence initial position might be falsely capitalized, we did some manual checking, but cannot claim to have achieved perfection.

Next, for each stimulus word, only the top five associations (i.e. the associations produced by the largest number of test person) were retained, and all other associations were discarded. The decision to keep only a small number of associations was motivated by the results of Rapp (2013) which indicate that associations produced by very few test persons tend to be of arbitrary nature. We also wanted to avoid unnecessary complications, which is why we decided on a fixed number, although the exact choice of five is of course somewhat arbitrary.

From the remaining dataset we removed all items which contained non-alphabetical characters. We also removed items which contained words that did not occur in the BNC. The reason for this is that quite a few of them are misspellings. By these measures, the number of items was reduced from initially 8400 to 7416.

From these we randomly selected 4000 items. 2000 of these were used as our training data set. The remaining 2000 were used as our test data set, but of course for the test set we removed the stimulus words. Tables 2 and 3 show the alphabetically first 20 items in each data set.[6]

The participating teams were asked to submit a list of 2000 words reflecting their predictions concerning the 2000 items of the test data set. For evaluation, we simply compared these 2000 words to the expected results (as taken from the EAT) by counting the number of exact matches, with the only flexibility that word capitalization was not taken into account.

There are a number of reasons why it was very difficult for the teams to get the target words exactly right:

1) In many cases, the given words might almost quite as strongly point to other target words. For example, when given the words *gin*, *drink*, *scotch*, *bottle*, and *soda*, instead of the target word *whisky* the alternative spelling *whiskey* should also be fine, and possibly some other beverages might also be acceptable.

2) The target vocabulary was not restricted in any way, so in principle hundred thousands of words had to be considered.

3) Although most of the target words were base forms, the training and the test sets also contain a good number of cases where the target words were inflected forms. Of course it is almost impossible to get these inflected forms exactly right.

Because of these difficulties we expected low performance figures (e.g. below 10%) in the competition[7] and were positively surprised by some of the actual results (see Section 5).

Concerning point 1 (other acceptable solutions) our data source did not provide any, so it was not practical for us to try to come up with alternative solutions in the chosen reverse association framework.

Concerning point 2 (restriction of target vocabulary), of course all teams had to make assumptions about the underlying vocabulary, as it is already difficult to fix boundaries for the English vocabulary, and occasionally even foreign words or names might occur as associations. In this respect all results have to be taken with caution, as some teams might have been more lucky than others in making good guesses concerning the target vocabulary.[8]

---

[5] Note that the participants of the shared task were nevertheless free to discard all case distinctions if their approach would not require them. During evaluation, case distinctions were not taken into account.

[6] From http://pageperso.lif.univ-mrs.fr/~michael.zock/CogALex-IV/cogalex-webpage/pst.html the full data sets can be downloaded

[7] Note that the results of up to 54% reported in Rapp (2014) were obtained using different data sets and severely restricted vocabularies, so these cannot be used for comparison.

[8] For such reasons we had requested to include such information in the papers. We concede that a competition with a pre-defined target vocabulary might have been more fair by reducing the influence of chance. But we were also very interested in the approaches on how to limit this vocabulary, so this was an important part of the shared task.

Target Word	Given Words
a	B the alphabet an man
abound	plenty many lots around leap
about	around turn round now time
above	below high over sky all
abrasive	rough sandpaper rub cutting hard
absence	away fonder illness leave presence
absent	away minded gone present ill
absurdity	stupid ridiculous mad stupidity clown
accents	dialects language foreign speech French
accordion	music piano play player instrument
accountant	money chartered clerk office turf
accrue	gather gain money acquire collect
achieve	nothing attain gain success win
acids	alkalis alkali bases burn science
acknowledged	letter receipt accepted received replied
acquaintance	friend know person friends casual
acquired	got obtained gained taste bought
acrid	smell bitter acid smoke dry
actions	words deeds movement movements reactions
actual	real fact happening truth exact

Table 2: Extract from the training set.

Given Words
able incapable brown clever good
able knowledge skill clever can
about near nearly almost roughly
above earth clouds God skies
above meditation crosses passes rises
abuse wrong bad destroy use
accusative calling case Latin nominative
ache courage blood stomach intestine
ache nail dentist pick paste
aches hurt agony stomach period
action arc knee reaction jerk
actor theatre door coach Act
actress stage play man theatre
addict pot store hash medicine
Africa Bible priest abroad doctor
again fresh afresh old morning
against angry bad fight hostile
age time epoch period years
aid assistant kind mother good
aid eyes aids see eye

Table 3: Extract from the test set. The respective (undisclosed) target words are shown in Table 4.

Concerning point 3 (matches of inflected forms) the ETS team had correctly pointed out that performance figures would significantly improve if matches with alternative inflected forms of the same word would also be counted as correct. For this purpose, the team kindly provided expanded versions of the target words for the training and for the test data set which were obtained using an in-house morphological tool. Table 4 shows the respective data for the alphabetically first 20 target words of the test data set. As we assumed that only the absolute but not the relative performance of the systems (ranking in competition) would be affected by this measure, we decided not to include this in the standard procedure, but nevertheless forwarded the data to all teams and encouraged them to conduct such an evaluation by themselves outside of the competition (and some actually did so). Let us nevertheless point out our main concerns:

1) Many target words are ambiguous, and in some cases the range of inflected forms depends on the way how the ambiguity is resolved. Assume, for example, that the target word form is *can* which might be an auxiliary verb or a noun. In this case, the inflected form *cans* in the expanded list would only be correct if the target word *can* referred to the noun, but not if it referred to the auxiliary verb (see also Lezius et al., 1998). Of course one could try to disambiguate the target words based on the given words. But this is a non trivial task likely to be error prone and possibly controversial.

2) In principle, such considerations might also apply to the given words, i.e. they could also be expanded. But in this case the disambiguation task is even more difficult as it is not clear what should be considered as context (i.e. as clues for disambiguation).

Although point 2 could be left to the participants, our aim was to avoid any such complications, in order to keep the focus on the core part of the shared task. So, as far as we as organizers were concerned, we decided not to consider inflectional variation.

Let us now comment on the overall character of the shared task. It should be noted that this task is actually the *reverse association task* as described in Rapp (2013, 2014). That is, the shared task participants were supposed to consider the associations from the EAT as their given words, and their task was to determine the original stimulus words.

Word	**Morphological expansions**
capable	
ability	abilities
approximately	
heavens	heaven
transcends	transcending, transcend, transcended
misuse	misusing, misused, misuses
vocative	vocatives
guts	gut, gutted, gutting
tooth	tooths
pains	pain, paining, pained
reflex	reflexes
stage	staging, staged, stages
actor	actors
drug	drugging, drugs, drugged
missionary	missionaries
anew	
antagonistic	
era	eras
helper	helpers
visual	visuals

Table 4: Morphological expansions of the first 20 words in the test data set.

However, we had not disclosed the nature of the data until after the competition mainly for the following reasons:

1) To avoid reverse engineering approaches based on the EAT or similar association norms.
2) To avoid leading participants in a particular direction. For us it seemed most important to obtain approaches as diverse as possible. And as this was the first shared task devoted to multi-stimulus associations, we thought that this would be a unique opportunity to obtain contributions as unbiased as possible.

On the other hand we had concerns about the fairness of not disclosing the nature of the data. Firstly, some of the participants might discover its origin and thus possibly have an advantage. Secondly, it is not clear in how far the reverse association task is prototypical enough for the lexical access problem as to assume that in terms of relative system performance the two tasks are comparable. In any case, concerning the lexical access problem we saw no chance of acquiring large scale data sets within the given time frame, so it was clear that this was not feasible.

When, after the competition, we disclosed the nature of the data, we invited the participants to comment on these issues in their papers, and it was very interesting for us to learn about the different views.

5 Participating systems and results

Altogether 15 teams expressed their interest to participate in the shared task. Of these, ten teams actually submitted results, of which one (BRNO) participated in both tracks (ukWaC and unrestricted), and another (SAAR) provided two solutions for the unrestricted track. The teams who submitted results are listed in Table 5, where each team is assigned a short Team ID which is derived from the institution names. In Table 6 for each team we make an attempt to give short characterizations of the approaches and the resources used.

Most approaches are variants of analyzing word co-occurrence statistics as derived from large text corpora. Several teams, among them the best performing ones, use for this purpose the open source tool *Word2Vec* which provides two neural network-based model architectures for computing continuous vector representations of words from very large data sets (Mikolov, 2013a; Mikolov, 2013b). In contrast, the RACAI team uses WordNet relation chains, a method which makes absolutely sense, but seems to severely suffer from data sparseness issues (i.e. there are much fewer WordNet relations between words than there are non-random word co-occurrences within large corpora). This finding is confirmed by the BRNO and UBC teams who tried out both approaches (corpus-based and WordNet-based) and came to the conclusion that the corpus-based approach performed considerably better.

Let us emphasize that we consider this type of findings a valuable output of the shared task and therefore are very grateful to the teams who pursued the WordNet-based approach that they shared these results although they were all well aware that, despite excellent scientific work, the respective performance figures were not very competitive.

Table 7 shows the results of the competition, ranked according to the accuracy of the results, and indicating the respective track (ukWAC or unrestricted). As some teams (AMU, QUT, SOEN, ranks 7 to 9) could not quite make it for the deadline, they were granted an extension of three days. On the top four positions are submissions who all used the above mentioned Word2Vec tool, indicating that this software is well suited for this task. Note that the winning system (IIITH) opted for the CBOW (continuous bag-of-words) architecture, whereas the other three opted for the skip-gram architecture. This might be an explanation for the differences in the results. However, this must be further analyzed as there are also other differences, including the assumptions constraining the target vocabulary, which, as described in Section 4, is an important issue. For example, the IIITH team used a frequency threshold of 25 while making word vectors using Word2Vec. In addition, when calculating PMI (pointwise mutual information) associations, a frequency threshold (for bigrams) of 3 was used (see sections 4.1 and 4.2 of their paper).

It should be mentioned that, like some others (see e.g. the papers by the ETS and by the RACAI teams), the IIITH team was able to improve on their results after the shared task deadline. Whereas for their submission they had used a re-ranking procedure based on point-wise mutual information (PMI), later on they used weighted PMI as their association measure. This improved their results from

30.45% to 34.9%. Likewise, the ETS team could improve their results from 14.95% to 18.90%. And the RACAI team (who used a WordNet-based approach) was able to almost double their results from 1.50% to 2.95%.

Team ID	Affiliation	Team members / Authors of papers
AMU	Aix-Marseille University, France	Gemma Bel-Enguix
BRNO	Brno University of Technology, Czech Republic	Lubomir Otrusina, Pavel Smrz
ETS	Educational Testing Service, Princeton, USA	Michael Flor, Beata Beigman Klebanov
IIIT	International Institute of Information Technology (IIIT), Hyderabad, India	Urmi Gosh, Sambhav Jain, Soma Paul
LEIPZIG	University of Leipzig, Germany	Rico Feist, Daniel Gerighausen, Manuel Konrad, Georg Richter, Thomas Eckart, Dirk Goldhahn, Uwe Quasthoff
QUT	Queensland University of Technology, Brisbane, Australia	Laurianne Sitbon, Lance De Vine
RACAI	Romanian Academy Research Institute for Artificial Intelligence, Bukarest, Romania	Catalin Mititelu, Verginica Barbu Mititelu
SAAR	Saarland University, Germany	Asad Sayeed (no paper)
SOEN	Universities of Stuttgart, Osnabrück, and Erlangen-Nürnberg, Germany	Gabrielle Lapesa, Stefan Evert
UBC	University of the Basque Country, Spain	Josu Goikoetxea, Eneko Agirre, Aitor Soroa

Table 5: Participating teams.

Team ID	Approach	Resources used
AMU	Co-occurrence-based lexical graph	British National Corpus
BRNO	Word2Vec from Python package GenSim (skip-gram architecture)	ukWaC, ClueWeb12, WordNet
ETS	Aggregating co-occurrence-based association strengths to individual cue words	English Gigaword 2003, ETS in-house corpus
IIITH	Word2Vec using CBOW architecture and re-ranking	ukWaC
LEIPZIG	Sum of co-occurrence-based significance values	Leipzig corpora collection
QUT	Own implementation similar to the Word2Vec package (skip-gram architecture)	ukWaC
RACAI	Shortest WordNet relations chain and maximum entropy modeling	Princeton WordNet, Google n-gram corpus
SAAR	Co-occurrence-based	ukWaC and others
SOEN	Ranking according to average (co-occurrence-based) association strength or according to distributional similarity	ukWaC
UBC	Word2Vec (skip-gram architecture), random walks, personalized PageRank	Google news corpus, Wikipedia, WordNet

Table 6: Overview on approaches and resources.

To give a rough idea on how much the results can be improved when inflectional variants are tolerated during evaluation (see Section 4), let us mention that the IIITH team did so. This way their results improved from 34.90% (as obtained after the deadline) to 39.55. Likewise, in the case of the ETS team the results improved from 14.95% to 20.25%. (For details see the respective contributions in these proceedings.)

Concerning the two tracks of the competition, namely ukWaC and unrestricted, it appears that the ukWaC corpus contains already enough information to solve the task. Evidence for this is provided by the BRNO team which submitted results in both tracks and where the improvements were minimal (19.85% vs. 19.65%). Another indication is that, unexpectedly, the winning IIITH team was in the ukWaC track.

For details on all other approaches (except SAAR) see the papers provided by the participating teams in these proceedings. Ideas that occurred when discussing the shared task with other colleagues were that Adam Kilgarriff's SketchEngine might be a useful tool for solving the lexical access problem (thanks to Eva Schaeffer-Lacroix for pointing this out), and that it may be useful to take syntax into account (thanks to Eric Wehrli and Luka Nerima). The latter would be in analogy to the generation of distributional thesauri where working with parsed rather than raw corpora has been shown to lead to very good quality (see e.g. Pantel & Lin, 2002). This way, rather than taking all word co-occurrences into account, the focus can be laid on selected relations between words, such as e.g. head-modifier or subject-object relations.

Rank	Team ID	Accuracy (%)	Track
1	IIITH	30.45	ukWAC
2	BRNO	19.85	unrestricted
3	BRNO	19.65	ukWaC
4	UBC	16.35	unrestricted
5	ETS	14.95	unrestricted
6	LEIPZIG	14.05	unrestricted
7	SOEN	13.10	ukWaC
8	AMU	9.10	unrestricted
9	QUT	4.25	ukWaC
10	SAAR	3.50	unrestricted
11	SAAR	2.60	unrestricted
12	RACAI	1.50	unrestricted

Table 7: Results of the shared task.

6 Discussion and conclusions

For the shared task of finding associations to multiple stimuli, by the participants accuracies of up to 30% (35% after the deadline) were reported. Given the very conservative evaluation procedure (see Section 4) which relies on exact matches and does not give any credit to alternative solutions, this is a very good result which considerably exceeded our expectations. Although we do not have comparative figures on human performance, our guess is that humans would not be able to do much better on this. So, in some sense, it seems that we have rather perfect results.

But what does this mean? Is there any psycholinguistic relevance? And is the task which we addressed here of any relevance for practical work in computational linguistics?

Let us first discuss the question of psycholinguistic relevance. In Rapp (2011) we have argued that human language intuitions are based on the detection, memorization, and reproduction of statistical regularities in perceived language. But we have only discussed this for single words. Now we can do so for multiword stimuli. And it seems that the same mechanisms that apply to single word stimuli are also valid in the case of multiwords. Apparently, from a relatively limited corpus such as ukWaC, intuitively plausible associations to an almost unlimited number of multiword stimuli can be derived. This is in analogy to human language acquisition: Due to limitations of the input channel a person can only perceive a few hundred million words during lifetime. But this limited information seems to suffice to have intuitions on almost anything that is language related.

This is a contradiction only on first glance: Apparently, language is a highly compressed form of information where all co-occurrences of words or word-sequences count (and were literally counted by most algorithms!). Therefore its information content is far higher than it may appear, and this provides a solution to the often discussed argument concerning the poverty of the stimulus (Landauer & Dumais, 1997). With regard to language, it seems there simply is no poverty of the stimulus, but instead the human language is a highly condensed form of extremely rich information. As the capacities of the input and the output channels are very limited, evolution was probably forced to optimize on this.

As the systems participating in the shared task can simulate human intuitions concerning zillions of possible multiword stimuli, it is likely that their algorithms grasp some of the essence that governs the respective inference processes taking place in human memory. In particular, they provide evidence that human association processing is also co-occurrence based, and that this not only applies to associations to single stimulus words as shown by Wettler et al. (2005), but also to associations concerning multiple stimuli.

Concerning the practical relevance of the work, our feeling is that such systems will be useful additions to many language-related tasks requiring human-like intuitions for the reason that human language intuitions seem to be based on associative learning. Let us come up with some examples of possible applications:

1) Augment associative resources such as the EAT.
2) Tip-of-the-tongue problem: Recall elusive words.
3) Lexical access: Rather than relying on alphabetical order, encyclopedias and dictionaries can be accessed associatively (e.g. *president of Poland → Bronislaw Komorowski*).
4) Generating thesauri of related words: Related words in the sense of Pantel & Lin (2002) are second order associations. The words related to a given word can be determined by computing its associations, and by then computing the multi-stimulus associations to these.
5) Question answering: Questions can be considered as multiword stimuli, answers as their associations (e.g. *height of Eiffel Tower → 324 m*).
6) Paraphrasing: The meaning of a phrase can be characterized by the associations resulting from its content words. Paraphrases are likely to lead to similar associations.
7) Search word generation in information retrieval: Keywords used in search queries can be augmented with relevant other keywords.
8) Advertising: The effect of an advertisement can be described by the associations evoked by the words that are used in it.
9) Word sense induction and disambiguation: Word contexts can be replaced by their multi-stimulus associations. This way the effects of word choice will be reduced when clustering contexts.
10) Machine translation: Translations can be seen as associations across languages (seed dictionary is required, see below).

Of course, most of the above has already been dealt with using other approaches. But, when looking at the respective (statistical) algorithms more closely, it seems often the case that researchers have intuitively chosen statistics which show some analogy to multi-stimulus associations. So what we suggest here is not entirely new. We nevertheless hope that the current framework can be useful. Firstly, it draws a connection to psycholinguistic evidence. And secondly, as done in the shared task, it allows to optimize the core algorithm independently of particular applications.

To be a bit more explicit, let us try to sketch a possible agenda of some future work which we would be happy to see: Let us start from the hypothesis that the meaning of a short sentence or phrase can be characterized by the vector resulting from taking its content words as multiword stimuli, and by computing their associations. For example, given the sentence *John laughed in the circus*, we would take *John*, *laugh*, and *circus* as our stimulus words, and the resulting association vector could be expected to have high values at its positions corresponding to *clown*, *nose*, and *fun*. For conciseness, let

us call this type of vector *meaning vector*.[9] Now let us look at the sentences *Someone walks across the border* and *A person passes customs*. The two sentences do not share a single word. But the associations derived from them should be nevertheless similar, because associations such as *toll*, *officer*, or *country* can be expected to come up in both cases. That is, their meaning vectors should be similar, and this similarity can be quantified e.g. by computing the cosine similarity between them. We thus have a method which allows us to measure the similarity between sentences in a way that to some extend takes their meanings into account.

Finally, we can try to cross language barriers and make the step to association-based machine translation (ABMT). To translate a source language phrase, we compute its meaning vector. Presupposing that we have a basic dictionary, in analogy to Rapp (1999) we can translate this meaning vector into the target language.[10] Further assuming that we already know the meaning vectors of a very large number of target language phrases, we next select the target language meaning vector which is most similar to the source language meaning vector. The respective target language phrase can be considered to be the translation of the source language phrase. Optionally, to improve translation quality, the target language phrase can be modified by adding, removing, substituting, or reordering words with the aim of improving the similarity between the meaning vectors of the source and target language phrases.

Acknowledgments

This work was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. We would like to thank George R. Kiss and colleagues for creating the Edinburgh Associative Thesaurus, and Michael Wilson for making it publicly available. Many thanks also to Adriano Ferraresi and colleagues for providing the ukWaC corpus, and to the participants of the shared task for their contributions and comments, as well as for the pleasant cooperation.

References

Aitchison, J. (2003). *Words in the Mind: an Introduction to the Mental Lexicon*. Oxford, Blackwell.

Bonin, P. (2004). *Mental Lexicon: Some Words to Talk about Words.* Nova Science Publishers.

Brown, A. (1991). A review of the *tip of the tongue* experience. *Psychological Bulletin*, 10, 204–223.

Brown, R. & Mc Neill, D. (1966). The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5: 325–337.

Burnard, L.; Aston, G. (1998): The BNC Handbook: Exploring the British National Corpus with Sara. Edinburgh: University Press.

Ceccato, S. (1956). La grammatiea insegnata alle machine. *Civiltà delle Machine*, Nos. 1 & 2.

Collins, A.M. & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* 8 (2): 240–247.

Collins, A.M. & Quillian, M.R. (1970). Does category size affect categorization time? *Journal of verbal learning and verbal behavior* 9 (4): 432–438.

Collins, A.M. & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review* 8.

Cramer, P. (1968). *Word Association*. Academic Press, New York.

Deese, J. (1965). The structure of associations in language and thought. Johns Hopkins Press. Baltimore

[9] As this is a bag-of-words approach which does not take syntax into account, of course we do not claim that such a vector can grasp all of a sentence's meaning.

[10] Note that gaps in dictionary coverage can be typically tolerated in such a setting as associations tend to be common words. That is, in principle the method allows to correctly translate words which are not in the dictionary. This is a property giving it some plausibility as a model for the cognitive processes underlying human translation.

Dutoit, D. and P. Nugues (2002): A lexical network and an algorithm to find words from definitions. In Frank van Harmelen (ed.): *ECAI2002, Proceedings of the 15th European Conference on Artificial Intelligence*, Lyon, 450–454.

Edmonds, D. (ed.), (1999). *The Oxford Reverse Dictionary*, Oxford University Press, Oxford, 1999.

El-Kahlout, I. D. and K. Oflazer. (2004). Use of Wordnet for Retrieving Words from Their Meanings. *Proceedings of the 2nd Global WordNet Conference*, Brno, 118–123.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database and some of its Applications. MIT Press.

Ferraresi, A.; Zanchetta, E.; Baroni M.; Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: S. Evert, A. Kilgarriff and S. Sharoff (eds.): *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakech.

Findler, N. (editor). (1979). Associative Networks: The Representation and Use of Knowledge by Computers. Academic Press, Inc., Orlando, FL, USA.

Fromkin V. (ed.). (1980). Errors in linguistic performance: Slips of the tongue, ear, pen and hand. New York: Academic Press.

Fromkin, V. (ed.) (1973): *Speech errors as linguistic evidence*. The Hague: Mouton Publishers

Kiss, G., Armstrong, C., Milroy, R. & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley and N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh: University Press.

Landauer, T.K.; Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104 (2), 211–240.

Lezius, W.; Rapp, R.; Wettler, M. (1998). A freely available morphology system, part-of-speech tagger, and context-sensitive lemmatizer for German. In: *Proceedings of COLING-ACL 1998,* Montreal, Vol. 2, 743–748.

Mandala, R., Tokunaga, T. & Tanaka, H. (1999). Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. *Proceedings of EACL.*

Mann, W. C. Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3), 243–281.

Meyer, D.E. & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90: 227–234.

Michiels, A. (1982). Exploiting a Large Dictionary Database. *PhD Thesis, University of Liège*, mimeographed.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Miller, G. A. (1995). WordNet : A lexical database for english. *Communications of the ACM*, 38 (11), 39–41.

Miller, G.A. (ed.) (1990): WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4), 235–244.

Nogier, J.F. & Zock, M. (1992) Lexical choice by pattern matching. *Knowledge Based Systems*, Vol. 5, No 3, Butterworth.

Pantel, P.; Lin, D. (2002): Discovering Word Senses from Text. *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02).* Edmonton, Canada , 613–619.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5), 410–430.

Quillian, M. R. (1968). Semantic memory. *Semantic Information Processing*, 227–270.

Quillian, M. R. (1969). The teachable language comprehender: a simulation program and theory of language. *Communications of the ACM*, 12(8), 459-476.

Quillian, R. (1963). A notation for representing conceptual information: An application to semantics and mechanical English paraphrasing. SP-1395, System Development Corporation, Santa Monica.

Quillian, R. (1966). *Semantic Memory*. Unpublished doctoral dissertation, Carnegie Institute of Technology.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics 1999,* College Park, Maryland. 519–526.

Rapp, R. (2011). Language acquisition as the detection, memorization, and reproduction of statistical regularities in perceived language. *Journal of Cognitive Science*, Vol. 12, No. 3, 297–322.

Rapp, R. (2013). From stimulus to associations and back. Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science, Marseille, France.

Rapp, R. (2014). Corpus-based computation of reverse associations. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Island.

Richens, R. H. (1956) Preprogramming for mechanical translation, *Mechanical Translation* 3 (1), 20–25.

Roget, P. (1852). Thesaurus of English Words and Phrases. Longman, London.

Schvaneveldt, R. (ed.) (1989). Pathfinder Associative Networks: studies in knowledge organization. Ablex. Norwood, New Jersey, US.

Simmons, R. (1963). Synthetic language behavior. *Data Processing Management* 5 (12): 11–18.

Sowa, John F. (1984). Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA.

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8, 418–425.

Summers, D. (1993). Language Activator: the world's first production dictionary. Longman, London.

Vitevitch, M. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research* , 51:408–422.

Wettler, M.; Rapp, R.; Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics* 12(2), 111–122.

Zock, M. (2014). How to overcome the tip-of-the-tongue problem with the help of a computer. Proceedings *of CogALex-IV*, COLING, Dublin, Ireland

Zock, M.; Cristea, D. (2014). You shall find the target via its companion words: specification of tools and resources to overcome the tip-of-the-tongue problem. *Proceedings of the 11th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, Venice.

Zock, M.; Ferret, O.; Schwab, D. (2010). Deliberate word access : an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, 13(4), 107–117.