Natural Language Engineering (2020), **26**, pp. 579–592
doi:10.1017/S1351324920000418

CAMBRIDGE
UNIVERSITY PRESS

EMERGING TRENDS

Benchmarks and goals

Kenneth Ward Church*

Baidu, Sunnyvale, CA 94089, USA
*Corresponding author. E-mail: kenneth.ward.church@gmail.com

(Received 25 June 2020; accepted 25 June 2020)

Abstract

Benchmarks can be a useful step toward the goals of the field (when the benchmark is on the critical path), as demonstrated by the GLUE benchmark, and deep nets such as BERT and ERNIE. The case for other benchmarks such as MUSE and WN18RR is less well established. Hopefully, these benchmarks are on a critical path toward progress on bilingual lexicon induction (BLI) and knowledge graph completion (KGC). Many KGC algorithms have been proposed such as Trans[DEHRM], but it remains to be seen how this work improves WordNet coverage. Given how much work is based on these benchmarks, the literature should have more to say than it does about the connection between benchmarks and goals. Is optimizing P@10 on WN18RR likely to produce more complete knowledge graphs? Is MUSE likely to improve Machine Translation?

Keywords: Benchmarks; Knowledge graph completion; Bilingual lexicon induction; MUSE; WordNet

1. Introduction

Many (perhaps most) papers in top conferences these days propose methods and test them on standard benchmarks such as General Language Understanding Evaluation (GLUE)[a] (Wang *et al.* 2018), Multilingual Unsupervised and Supervised Embeddings (MUSE)[b] (Conneau *et al.* 2017), and wordnet 18 reduced relations (WN18RR) (Dettmers *et al.* 2018). Some of these methods (Bidirectional Encoder Representations from Transformers (BERT) Devlin *et al.* 2019, enhanced representation through knowledge integration (ERINE) Sun *et al.* 2020) not only do well on benchmarks but also do well on tasks that we care about.[c] Unfortunately, despite large numbers of papers with promising performance on benchmarks, there is remarkably little evidence of generalizations beyond benchmarks.

In my last Emerging Trends column (Church 2020), I complained about reviewing. Reviewers do what reviewers do. They love papers that are easy to review. Reviewers give high grades to boring incremental papers that do slightly better than SOTA (state of the art) on an established benchmark. No one ever questions whether the benchmark is still relevant (or whether it was ever relevant). Generalizing beyond the benchmark is of little concern. Precedent is good (and simple to review). Impact is too much work for the reviewers to think about.

Benchmarks have a way of taking on a life of their own. Benchmarks tend to evolve over time. When the benchmark is first proposed, there is often a plausible connection between the

[a] https://super.gluebenchmark.com/.
[b] https://github.com/facebookresearch/MUSE.
[c] https://www.blog.google/products/search/search-language-understanding-bert/.

benchmark and a reasonable goal that is larger than the benchmark. But this history is quickly forgotten as attention moves to SOTA numbers, and away from sensible (credible and worthwhile) motivations.

2. The goal

Before discussing some of the history behind MUSE and WN18RR, benchmarks for bilingual lexicon induction (BLI) and knowledge graph completion (KGC), it is useful to say a few words about goals. It is important to remember where we have been, but even more important to be clear about where we want to go.

I do not normally have much patience for management books, but “The Goal” (Goldratt 1984) is an exception.[d] The point is that one should focus on what matters and avoid misleading internal metrics. Goldratt invents a fictional factory to make his point. The fictional factory is losing money because of misleading internal metrics. They should be focused on end-to-end profit, a combination of three factors: (1) throughput (sales), (2) inventory costs, and (3) operational expense. But in the story, they introduced a misleading metric (output per hour), which appeared to be moving in the right direction when they introduced automation (robots), but in fact, the robots were increasing cost, because most of the output was ending up in inventory. The moral of the story is that we need to focus on what matters (end-to-end results). It can be helpful to factor a complicated system into simpler units that are easier to work with (unit testing is often easier and more actionable than system testing), but we need to make sure the simplification is on the critical path toward the goal.

Another example[e] is a jewelry business. Unlike the fictional factory, this is a real example. This jewelry business has lots of products (and consequently, lots of inventory). Products have a long tail. Some products sell faster than others. The right metrics helped the business focus on key bottlenecks. Two simple process improvements increased sales:

1. The business had been prioritizing too many products, but better metrics encouraged them to prioritize products by sales. Make sure that shelves are stocked with fast movers, before stocking shelves with other products.
2. In addition, if a product has been in a store for a while and has not sold, rotate it to another store. Some products sell better in some markets and other products sell better in other markets.

The moral, again, is to focus on the right metrics. Without focus, there is a tendency to optimize everything. But most systems are constrained by a few bottlenecks. Optimize bottlenecks (and nothing else). Misleading metrics are worse than useless because of opportunity costs. The wrong metrics distract attention from what matters (addressing bottlenecks) and encourage wasted effort optimizing steps that are not on the critical path toward the goal. Resist the temptation to optimize everything (because most things are not on the critical path toward the goal).

What does this have to do with benchmarks? Benchmarks can be a useful step toward the goal (when the benchmark is on the critical path), as demonstrated by the GLUE benchmark, and deep nets such as BERT and ERNIE. The case for other benchmarks such as MUSE and WN18RR is less well established. Survey articles (Nguyen 2017) mention many KGC algorithms: Trans[DEHRM], KG2E, ConvE, Complex, DistMult, and more (Bordes *et al.* 2013; Wang *et al.* 2014; Yang *et al.* 2015; Lin *et al.* 2015; Trouillon *et al.* 2016; Nguyen *et al.* 2017; Nickel, Rosasco, and Poggio 2019; Sun *et al.* 2019). Many of these perform well on the benchmark, but it remains to be seen how this work improves coverage of WordNet (Miller 1998). Will optimizing P@10 on WN18RR produce more complete knowledge graphs? Has WordNet coverage improved as a result of all this work on the KGC WN18RR benchmark? If not, why not? When should we expect to see such results?

[d] https://www.youtube.com/watch?v=2RVMgV37O_k.
[e] https://youtu.be/_COdSwmIDMY?t=641.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

3. Background: Rotation matrices, BLI and KGC

Rotation matrices play an important role in both BLI and KGC. Benchmarks in both cases provide various resources (embeddings Mikolov, Le, and Sutskever 2013; Pennington, Socher, and Manning 2014; Mikolov *et al.* 2017) as well as test and train sets. In particular, MUSE[f] provides embeddings in 45 languages, as well as training and test dictionaries (in both directions) between English (en) and 44 other languages: af, ar, bg, bn, bs, ca, cs, da, de, el, es, et, fa, fi, fr, he, hi, hr, hu, id, it, ja, ko, lt, lv, mk, ms, nl, no, pl, pt, ro, ru, sk, sl, sq, sv, ta, th, tl, tr, uk, vi, zh. In addition, MUSE provides bilingual dictionaries for all pairs of six languages: en, de, es, fr, it, pt. (This paper will use ISO 639-1 for languages).[g]

The training dictionaries are also known as seed dictionaries. Seed dictionaries, S, consist of $|S|$ pairs of translation equivalents, $< x_i, y_j >$, where x_i is a word in source language x and y_j is a word in target language y. These dictionaries are used to train a rotation. That is, we construct two arrays, X and Y, both with dimensions $|S| \times K$. Rows of X are $vec(E_x, x_i)$ and rows of Y are $vec(E_y, y_i)$, where $vec(E_x, x_i)$ looks up the word x_i in the embedding, E_x for language x. Embeddings, E_x and E_y have dimensions, $V_x \times K$ and $V_y \times K$, respectively, where V_x and V_y are the sizes of the vocabularies for the two languages.

The training process solves the objective:

$$\min_{R_{x,y}} \|XR_{x,y} - Y\|_F^2 \quad \text{where} \quad R_{x,y} \in \mathcal{R}^{K \times K} \tag{1}$$

$R_{x,y}$ is a $K \times K$ rotation matrix that translates vectors in language x to vectors in language y. A simple solution for R is the Orthogonal Procrustes problem.[h]

At inference time, we are given a new source word, x_i in the source language x. The task is to infer $trans_{x,y}(x_i) = y_j$, where y_i is the appropriate translation in target language y. The standard method is to use Equation (2), where vec^{-1} is the inverse of vec. That is, vec^{-1} looks up a vector in an embedding and returns the closest word.

$$trans_{x,y}(x_i) \approx vec^{-1}(E_y,\ vec(E_x, x_i)\, R_{x,y}) \tag{2}$$

Much of the BLI literature improves on this method by taking advantage of constraints such as hubness (Smith *et al.* 2017). Most words have relatively few translations, and therefore, the system should learn a bilingual lexicon with relatively small fan-in and fan-out. Hubs are undesirable; we do not want one word in one language to translate to too many words in the other language (and vice versa).

Consider random walks over MUSE dictionaries. If we start with *bank* in English, we can translate that to *banco* and *banca* in Spanish. From there, we can get back to *bank* in English, as well as *bench*.

$$bank \rightarrow banco|banca \rightarrow bank|bench \tag{3}$$

Table 1 shows that most words (in the MUSE challenge) have very limited fan-out. In fact, most words are disconnected islands, meaning they back-translate to themselves and nothing else (via random walks over 45 language pairs).

Taking advantage of hubness clearly improves performance on the MUSE challenge, but why? Hopefully, the explanation is the one above (most words have relatively few translations), but it is also possible that hubness is taking advantage of flaws in the benchmark such as gaps in MUSE (most words should have many more translations than those in MUSE (Kementchedjhieva, Hartmann, and Søgaard 2019)).

[f] https://github.com/facebookresearch/MUSE.
[g] https://docs.oracle.com/cd/E13214_01/wli/docs92/xref/xqisocodes.html.
[h] https://en.wikipedia.org/wiki/Orthogonal_Procrustes_problem.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 1. Fan-out for 168k English words in MUSE, most of which (115k) are disconnected islands that back-translate to themselves (and nothing else). The majority of the rest back-translate to five or more words

0	1	2	3	4	5+
1209	115,178	4172	1269	540	45,587

3.1 Knowledge Graph Completion (KGC)

KGC benchmarks (WN18RR) are similar to BLI mechmarks (MUSE). Many of these KGC algorithms (and evaluation sets) are now available in pykg2vec (Yu *et al.* 2019), a convenient Python package.[i] The goal of KCG, presumably, is to improve coverage of knowledge graphs such as WordNet.

KGC starts with $< h, r, t >$ triples, where h (head) and t (tail) are entities (words, lemmas, or synsets) connected by a relation r. For example, the antonymy relationship $<$ *inexperienced*, $\neq$, *experienced*$>$, is a triple where h is *inexperienced*, r is $\neq$ and t is *experienced*. Heads and tails are typically represented as vectors, and relations are represented as rotation matrices. Thus, for example, antonymy could be modeled as a regression task: $vec(inexperienced) \sim vec(experienced)$, where the slope of this regression is a rotation matrix. In this case, the rotation matrix is an approximation of the meaning of negation.

Benchmarks such as WN18RR are incomplete subsets of WordNet. WN18RR consists of 41k entities (WordNet synsets) and 11 relations, though just 2 of the relations cover most of the test set. We can model triples as graphs, $G_r = (V, E)$, one for each relation r, where V is a set of vertices (h and t), and E is a set of edges connecting h to t. WN18RR splits edges, E, into train, validation and test randomly, with 60% in train, and 20% in each of the other two sets. At training time, we learn a model from the training set. At inference time, we apply that model to a query from the test set, $< h, r, ? >$ or $<?, r, t >$. The task is to fill in the missing value. N-best candidates are scored by precision at ten (P@10).

This setup is more meaningful when edges are iid, though WordNet makes considerable use of equivalence relations (synonyms), apartness relations (antonyms), partial orders (is-a, part-of), and other structures that are far from iid.

Unstructured: Edges are iid. If we tell you there is (or is not) an edge between h and t, we have provided no information about the rest of the graph.

Structured: Edges are not iid. Examples: equivalence classes and partial orders.

WN18RR is an improvement over an earlier benchmark, WN18, which suffered from information leakage (Dettmers *et al.* 2018). WordNet documentation[j] makes it clear that various links come in pairs (by construction). If a *car* is a *vehicle*, for example, then there will be both a hypernym link in one direction, as well as a hyponym link in the other direction. Similar comments apply to other relations such as part-of. WN18 originally had 18 relations, but the 18 were reduced down to 11 in WN18RR.

WN18RR addresses some of the leakage, but there is more. Most of the test set is dominated by 2 of the 11 relations, hyperym and derivationally related forms. The former is a partial order (is-a) and the latter is (nearly) an equivalence relation, combining aspects of morphology with synonymy. We recently submitted a paper questioning the use of iid assumptions for equivalence relations. It turns out that one can do very well on the benchmark (without addressing the goal of improving WordNet coverage) because much of the test set can be inferred from random walks (and transitivity) from triples in the training set.

[i] https://github.com/Sujit-O/pykg2vec.

[j] https://globalwordnet.github.io/gwadoc/.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 2. Five synsets in two (of 29) languages

Synset	en	fr
bank.n.01	bank	banque, rive
bank.n.03	bank	banque
bank.n.04	bank	**NA**
bank.n.05	bank	banque
bank.n.06	bank	banque, rive

4. WordNet is incomplete because it is too English-centric

The missing at random model does not pay enough respect to Miller, an impressive researcher (and Chomsky's mentor). One is unlikely to improve his work much by the kinds of methods that have been discussed thus far such as rotation matrices and random walks. To make meaningful progress, we need to bring something to the table (such as more languages) that goes well beyond the topics that Miller was thinking about.

Miller was focused on English. WordNets are now available in German (Hamp and Helmut 1997), Chinese (Dong, Dong, and Hao 2010), and many other languages.[k] The Natural Language Toolkit (NLTK) interface to WordNet[l] supports 29 languages: ar, bg, ca, da, el, en, es, eu, fa, fi, fr, gl, he, hr, id, it, jp, ms, nb, nl, nn, pl, pt, qc, sl, sq, sv, th, zh.[m] Coverage varies considerably by language. Few languages have as much coverage as English. Some have considerably less.

Table 2 illustrates glosses in English (en) and French (fr) for five synsets of bank. All five synsets have a single gloss in English. Two synsets have two glosses in French, and one has none.

The basic framework was developed for English. As WordNet becomes more and more multilingual, it needs to move away from viewing English as a pivot language toward a more language universal inter-lingua view of world knowledge. Ultimately, the knowledge completion goal should aim higher than merely capturing what is already in WordNet (English-centric knowledge) to something larger (world knowledge). Benchmarks such as WN18RR distract us away from the larger goal (capturing world knowledge that goes beyond the English-centric view that Miller was working with), to something smaller than what Miller was thinking about (how to capture subsets of English from other subsets of English).

Other languages bring new insights to the table. We recently submitted a paper proposing a new similarity metric, backsim (back-translation similarity). Backsim uses bilingual dictionaries in MUSE to find 316k pairs of words like *similarly* and *likewise* that have similar translations in other languages. We suggested that many (271k) of these pairs should be added to WordNet under four relations: M (morph), S (synonym), H (hypernym), and D (derived form). In this way, other languages help us move from our own view of our language toward world knowledge.

4.1 Comparisons of WordNet and MUSE

We have an embarrassment of riches now that WordNet is available in 29 languages, and MUSE is available in 45 languages. Comparisons of the two (Tables 3–5) suggest WordNet has remarkably

[k] https://globalwordnet.org/resources/wordnets-in-the-world/.

[l] https://www.nltk.org/howto/wordnet.html.

[m] We use 2-letter codes for languages (ISO639-2); WordNet uses 3-letter codes. Two of the 29 languages are not defined in www.loc.gov/standards/iso639-2/php/code_list.php: zsm, qcn. We use ms for zsm.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 3. Vocabulary Sizes (excluding disconnected islands). Numbers are larger for WordNet than MUSE, suggesting WordNet has better coverage. Numbers are also larger for English (en) than other languages, suggesting both WordNet and MUSE have better coverage of English than other languages

Source	Pivot language (MUSE)				
Language	en	es	fr	it	pt
en		19,296	20,159	14,700	15,478
es	23,335		588	8173	9422
fr	22,727	1056		9465	11,549
it	17,948	1697	1161		942
pt	21,418	1699	4306	721	
Source	Pivot language (WordNet)				
Language	en	es	fr	it	pt
en		46,574	77,145	46,041	58,920
es	24,627		18,499	11,157	13,670
fr	39,946	20,359		22,715	30,920
it	33,225	17,334	28,056		22,888
pt	41,743	21,995	37,234	24,242	

good coverage, probably better than MUSE. Researchers in BLI would be well advised to consider WordNet in addition to (or perhaps as a substitute for) the bilingual dictionaries in MUSE.

Table 3 compares MUSE and WordNet vocabulary sizes after removing disconnected islands. Let $M_{x,y}$ be a sparse matrix with a non-zero value in cell i, j if there is a translation from x_i to y_j, where x_i is a word in language x and y_j is a word in language y. We then form $M_{x,x} = M_{x,y}M_{y,x}$. This matrix, $M_{x,x}$, tells us how words in language x can back-translate via the pivot language y. Most words back-translate to themselves and nothing else. We refer to these words as disconnected islands. We are more interested in words with more translation possibilities.

Table 3 suggests WordNet has better coverage than MUSE since numbers are larger for WordNet than MUSE. In addition, the table suggests that both WordNet and and MUSE have better coverage of English (en) than other languages.

Some of the numbers in Table 3 are embarrassingly small, especially for MUSE. Numbers under 10k are suspiciously low. In MUSE, some languages have only 1k words with more than one back translation ($\approx$ 1% of the total vocabulary). In other words, some of the MUSE dictionaries are close to a substitution cipher. This may explain why hubness has been so effective for MUSE. But if this is the explanation, it may cast doubt on how effective hubness methods will be for generalizing beyond the benchmark.

Tables 4–5. dive deeper into French glosses. These tables show the number of glosses for nearly 18k English words that have at least 1 gloss in both collections. Of these, there are 10,330 words with more French glosses in WordNet, and 2266 with more in MUSE, and 5125 with the same number of French glosses in both collections.

In addition to simple counts, we would like to know if we are covering the relevant distinctions. One motivation for word senses is translation. Bar-Hillel (1960) left the field of machine translation because he could not see how to make progress on word sense disambiguation (WSD). The availability of parallel corpora in the early 1990s created an opportunity to make progress on

Table 4. WordNet (WN) has more French glosses than MUSE

Word	WN	MUSE
resolve	14	2
peel	7	2
recommend	4	2
ortolan	4	1
genre	3	2
transducer	2	1
leper	2	1
celibacy	2	1
armory	2	1
rind	1	1

Table 5. WordNet has more French glosses than MUSE. Each cell, *i*, *j*, counts the number of words with *i* glosses in WordNet and *j* glosses in MUSE

	Glosses in MUSE					
Glosses in WordNet	1	2	3	4	5	6
1	3880	1021	304	163	50	5
2	2305	877	342	151	40	2
3	1394	652	266	118	27	4
4	854	440	187	83	36	2
5	578	331	152	66	19	1
6	326	198	119	62	28	0
7	269	165	90	43	12	0
8	175	135	64	36	17	1
9	129	85	50	32	6	0
10+	381	384	297	195	68	4

Bar-Hillel's concerns. WSD could be treated as a supervised machine learning problem because of an interaction of word senses and glosses, as illustrated in Table 6 (Gale, Church, and Yarosky 1992).

How effective are WordNet and MUSE on examples such as those in Table 6? Unfortunately, WordNet glosses both bank.n.01 and bank.n.06 with *banque*, even though the definitions and examples make it clear that bank.n.01 is a river bank, and bank.n.06 is a money bank. Similarly, WordNet glosses for *drug* mention *drogue* but not *médicament*, possibly because this legal/illegal distinction is less salient in English. It is well known in lexicography that monolingual concerns are different from bilingual concerns. One should not expect a taxonomy of English synsets to

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 6. Interaction between word sense (English) and glosses (French)

English	Sense	French	Sense	French
bank	money	banque	river	banc
drug	illegal	drogue	legal	médicament
sentence	judicial	peine	grammatical	phrase
duty	tax	droit	obligation	devoir
land	property	terre	country	pays
language	medium	langue	style	langage
position	place	position	job	poste

generalize very well to all the worlds' languages. Thus, we see gaps across languages as a better opportunity to improve WordNet than gaps within English. Miller already thought quite a bit about coverage within English, but not so much about coverage across languages.

In short, spot checks are not encouraging; the coverage of glosses in WordNet and MUSE are probably inadequate for WSD applications, at least in the short term, where other approaches to WSD are more promising. Longer term, we see WSD applications, especially in a bilingual context, as an opportunity to test KGC, with less risk of leakage than benchmarks such as WN18RR. That is, the task is not to predict held-out edges but to predict translations of polysemous words into other languages. There is no shortage of testing and training material for this task, given how much text is translated, and how many of those words are polysemous. For a task like this, one might expect methods based on Machine Translation (Wu *et al.* 2016) and/or BERT/ERNIE Transformers (Devlin *et al.* 2019; Sun *et al.* 2020) to offer stronger baselines than traditional KGC methods.

4.2 WordNet as a database: Unifying BLI and KGC

Thus far, we have been treating bilingual lexicons and knowledge representation as separate problems, but there are aspects of both in WordNet, which become more salient when we view WordNet as a simple database. The proposed database view factors some facts into tables that depend on language, and other facts into tables that are language universal.

This database view of WordNet combines various aspects of both MUSE and WN18RR. WordNet has glosses over multiple languages (like MUSE), as well as relations over entities (like WN18RR). The schema is a bit more complicated than both MUSE and WN18RR because there are multiple tables (some depend on language and some do not), and three types of entities: synsets, lemmas, and strings (glosses, examples, and definitions).

1. Synset relations: $< h, r, t >$ where h (head) and t (tail) are synsets and r is a member of a short list of relations on synsets (e.g., is-a, part-of).
2. Definitions: maps 118k synsets to definitions (English strings).
3. Examples: maps 33k (of 118k) synsets to examples (English strings).
4. Synset2lemmas: maps synset to 0 or more lemmas in 29 languages
5. Lemma Glosses: maps lemma objects to glosses (strings) in 29 languages
6. Lemma Relations: $< h, r, t, lang >$, where h and t are lemmas and r is a member of a short list of relations on lemmas (e.g., derivationally related forms, synonyms, antonyms, and pertainyms), and *lang* is one of 29 languages.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 7. Sizes of WordNet tables

Table	Rows	Schema	Universal?
Synset Relations	198k	$< head, rel, tail >$	Universal
Definitions	118k	$synset \rightarrow def(str)$	Currently, English
Examples	33k	$synset \rightarrow \{example(str)\}$	Currently, English
Synset2lemmas	118k	$synset \times 29\ lang \rightarrow \{lemma\}$	Language Specific
Lemma Glosses	459k	$lemma \times 29\ lang \rightarrow \{gloss(str)\}$	Language Specific
Lemma Relations	2.9M	$< head, rel, tail, lang >$	Language Specific

Sizes of these tables are reported in Table 7. This view was extracted using a very simple program based on the NLTK interface to WordNet.[n]

It would be worthwhile to move some relations from the language-specific Lemma Relations table to the language universal Synset Relations table, but doing so would require refactoring WordNet in ways that probably require considerable effort. For example, there are many more antonyms in English than in other languages, not because English has more antonyms, but because the community has not yet made the effort to port antonym relations to other languages. It would be even better, perhaps, to move antonym relations from the language-specific Lemma Relations table to the language universal Synset Relations table, but that is likely to be a substantial undertaking.

In short, while WordNet is far from complete, and remains, very much, a work in progress, it is, nevertheless, an amazing resource. In comparison to benchmarks such as MUSE and WN18RR, WordNet has a number of advantages. In addition to coverage, the schema reflects that fact that considerable thought went into WordNet. WordNet is not only Miller's last (and perhaps greatest) accomplishment, but it has also benefited by years of hard work by a massive team working around the world in many languages. There are undoubtedly ways for machine learning to contribute, but such contributions are unlikely to involve simple techniques such as rotations and transitivity. There are likely to be more opportunities for machine learning to capture generalizations across languages than within English because Miller was thinking more about English and less about other languages.

5. History and motivation for BLI and MUSE benchmark

The previous section proposed a unified view of BLI and KGC, where WordNet can be viewed as combining aspects of both literatures. Obviously, the two literatures have quite different histories.

BLI has received considerable attention in recent years (Mikolov *et al.* 2013; Irvine and Callison-Burch 2013; Irvine and Callison-Burch 2017; Ruder, Vulić, and Søgaard 2017; Artetxe, Labaka, and Agirre 2018; Huang, Qiu, and Church 2019), though the idea is far from new (Rapp 1995; Fung 1998). BLI starts with comparable corpora (similar domains, but different language and different content), which are easier to come by than parallel corpora (literal sentence-for-sentence translations of the same content).

BLI is not on the critical path toward the goal when there are other methods that work better. When we have parallel corpora, we should use them. The technology for parallel corpora (Brown *et al.* 1993; Koehn *et al.* 2007) is better understood and more effective than the technology for

[n] https://github.com/kwchurch/Wordnet_tables.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

comparable corpora. The opportunity for comparable corpora and BLI methods should be a Plan B. Plan A is to get what we can from parallel corpora, and Plan B is to get more from comparable corpora.

Much of the recent BLI work focuses on general vocabulary, but the big opportunity for BLI is probably elsewhere. It is unlikely that BLI will be successful going head-to-head with parallel corpora on what they do best. Hopefully, terminology is a better opportunity for BLI.

Comparable corpora were proposed in the 1990s, as interest in parallel corpora was beginning to take off. It was clear, even then, that availability of parallel corpora would be limited to unbalanced collections such as parliamentary debates, (Canadian Hansards[o] United Nations,[p] Europarl Koehn 2005). Lexicographers believe that balance is very important, as discussed in Section 6.1 of Church and Mercer (1993). Fung proposed comparable corpora to address concerns with balance. She realized early on that it will be easier to collect balanced comparable corpora than balanced parallel corpora.

If one wants to translate medical terms such as MeSH,[q] parliamentary debates are unlikely to be helpful. It is better to start with corpora that are rich in medical terminology such as medical journals. There are some small sources of parallel data such as the New England Journal of Medicine (NEJM),[r] and much larger sources of monolingual data such as PubMed.[s] At Baidu, we can find some comparable monolingual data in Chinese (though it is difficult to share that data).

We would love to use BLI methods to translate the more difficult terms in PubMed abstracts. Obviously, many of these terms are not well covered in parallel corpora mentioned above (NEJM is too small, and parliamentary debates are too irrelevant). Unfortunately, technical terms are challenging for SOTA BLI methods, because BLI methods are more effective for more frequent words (and technical terminology tends to be less frequent than general vocabulary, even in medical abstracts).

Lexicographers distinguish general vocabulary from technical terminology. Dictionaries focus on general vocabulary, the words that speakers of the language are expected to know. Dictionaries avoid technical terms, because there are too many technical terms (too much inventory), and the target market of domain experts is too small (insufficient sales). Only a relatively small set of domain experts care about technical terms, but everyone cares about general vocabulary. The marketing department can reasonably plan on selling a dictionary on general vocabulary to a large market; it is harder to make the business case work for specialized vocabulary given the smaller market (and larger inventory costs). If BLI could make a serious dent on inventory costs, that might change the business case.

There is a need for a solution for specialized vocabulary. People are not very good at translating terminology. Professional translators live in fear of terminology. Everyone in the audience knows the subject better than the translators. Translators would rather not make it clear to the audience that they do not know what they are talking about. Terminology mistakes are worse than typos. With a typo, there is a possibility that the author knew how to spell the word but failed to do so. On the other hand, terminology mistakes make it clear to all that the translator is unqualified in the subject matter.

In computational linguistics, there is a tendency to gloss over the difference between specialized vocabulary and technical terminology. Benchmarks like MUSE make it easy to view the BLI task as a simple machine learning task, with no need to distinguish specialized vocabulary from general vocabulary. But students that study in America have more appreciation for the translators' predicament. These students know they cannot give their job talk in their first language because they do not know the terminology in their first language.

[o] https://catalog.ldc.upenn.edu/LDC95T20.
[p] https://catalog.ldc.upenn.edu/LDC94T4A.
[q] https://www.nlm.nih.gov/mesh/meshhome.html.
[r] https://github.com/boxiangliu/med_translation.
[s] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Table 8. MLI (monolingual lexicon induction) results for challenging test set of 10k PubMed terms (ranks 50–60k). P@1 is disappointing when embeddings are trained on relevant data (and worse when trained on irrelevant data)

P@1	Mean Score1	
PubMed	**0.425**	**0.49**
Crawl	0.088	0.36
Wiki	0.039	0.33
GNews	0.016	0.34

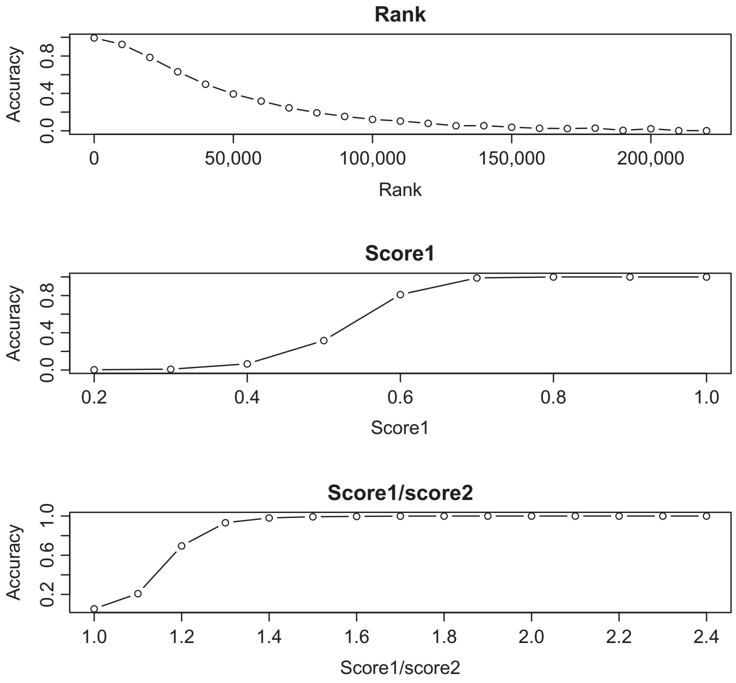

Figure 1. BLI technology is more effective for high-frequency words. Accuracy is better for high rank (top), large $score_1$ (middle) and large gap between $score_1$ and $score_2$ (bottom).

Current BLI technology may be effective for general vocabulary (the top 50k words), but BLI is less likely to be effective for relatively infrequent terminology as shown in Table 8. None of the P@1 scores in Table 8 are encouraging. P@1 is better when embeddings are trained on relevant data than irrelevant data, but P@1 is not particularly encouraging for difficult terms, even when trained on relevant data.

Table 8 and Figure 1 are based on a method we call monolingual lexicon induction (MLI). MLI is like BLI except that both the source and target embeddings are in the same language. In this case, the two embeddings were trained on two samples of PubMed abstracts (in English). The task is to learn the identity function. Can BLI methods discover that technical terms translate to themselves (when the source and target language are the same)? If not, BLI methods are even less likely to work when the source and target language are different. The plots in Figure 1 are smoothed with a simple logistic regression model for clarity.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

The main point of Figure 1 is the decline of BLI effectiveness with rank. BLI is relatively effective for low-rank (high-frequency) words, but less effective for high-rank (low-frequency) words. This may well be a fundamental problem for BLI. There may not be a sweet spot. BLI may not be on the critical path toward any goal. For high-frequency words (general vocabulary), there are better alternatives (parallel corpora), and for low-frequency words (specialized vocabulary), BLI is relatively ineffective.

Figure 1 makes a couple of additional points: BLI is more effective when (a) $score_1$ is large and (b) $score_2$ is not.

1. $score_1$: cosine of query term and top candidate (appropriately rotated), and
2. $score_2$: cosine of query and next best candidate.

This might suggest a more promising way forward toward a reasonable goal. There is hope that we could use features such as $score_1$ and $score_2$ to know which terms are likely to be translated correctly and which are not. BLI might be useful, even with fairly low accuracy, if we knew when it is likely to work, and when it is not. Unfortunately, benchmarks such as MUSE encourage optimizations that are not on the critical path toward a reasonable goal and discourage work on more promising tasks such as reject modeling.

Currently, machines are better than people for some tasks (e.g., spelling), and people are better than machines for other tasks (e.g., creative writing). Machines have an unfair advantage over people with spelling because machines can process more text. If BLI really worked, then terminology would be more like spelling than creative writing.

To conclude, comparable corpora were introduced almost 30 years ago to address lexicographers' concern with balance and translators' concerns with terminology. Since then, the emphasis has moved onto benchmarks and SOTA numbers. But this emphasis on numbers is probably not addressing realistic goals such as the original motivations: balance and terminology.

BLI should not compete with parallel corpora on general vocabulary, where BLI is not on the critical path (because parallel corpora work better than comparable corpora for general vocabulary). BLI could be useful, if it returned to the original motivations. In particular, there may be opportunities for BLI technology to play a role in terminology, especially in relatively modest glossary construction tools (Dagan and Church 1994; Justeson and Katz 1995; Smadja, McKeown, and Hatzivassiloglou 1996; Kilgariff *et al.* 2004).

6. Conclusions

Benchmarks have a way of taking on a life of their own. We have a tendency to take our benchmarks too seriously. Feynman warned us not to fool ourselves:

The first principle is that you must not fool yourself—and you are the easiest person to fool

Benchmarks can be like the misleading internal metrics that Goldratt warned us about. Given how much work has gone into developing methods based on the benchmarks, there should be more concern in the literature about goals. Why do we believe that optimizing scores on MUSE and WN18RR will bring us closer to larger goals like BLI and KGC? Is there evidence that systems that do well on these benchmarks generalize to problems that we care about?

References

Artetxe M., Labaka G. and Agirre E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*.

Bar-Hillel Y. (1960). The present status of automatic translation of languages. *Advances in Computers* **1**(1), 91–163.

Bordes A., Usunier N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, pp. 2787–2795.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Brown P., Della Pietra S.A., Della Pietra V.J. and Mercer R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2), 263–311, MIT Press.

Conneau A., Lample G., Ranzato M., Denoyer L. and Jégou H. (2017). Word translation without parallel data. arXiv preprint arXiv:1710.04087.

Church K. and Mercer R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* **19**(1), 1–24.

Church K. (2020). Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering* **26**(2), 245–257, Cambridge University Press.

Dagan I. and Church K. (1994). Termight: Identifying and translating technical terminology. In *Applied ACL*, pp. 34–40.

Dettmers T., Minervini P., Stenetorp P. and Riedel S. (2018). Convolutional 2D knowledge graph embeddings. In *AAAI*.

Devlin J., Chang, M.-W., Lee K. and Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186.

Dong Z., Dong Q. and Hao C. (2010). *HowNet and the Computation of Meaning*, Coling, pp. 53–56.

Fung P. (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Conference of the Association for Machine Translation in the Americas*. Springer, pp. 1–17.

Gale W., Church K. and Yarosky D. (1992). *Computers and the Humanities* **26**(5–6), 415–439, Springer.

Goldratt E.M. and Cox J. (1984). *The Goal: A Process of Ongoing Improvement*, Routledge, UK.

Hamp B. and Feldweg H. (1997). Germanet – A lexical-semantic net for german. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, ACL Workshop*.

Huang J., Qiu Q. and Church K. (2019). Hubless nearest neighbor search for bilingual lexicon induction. In *ACL*, pp. 4072–4080.

Irvine A. and Callison-Burch C. (2017). A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics* **43**(2), 273–310, MIT Press.

Irvine A. and Callison-Burch C. (2013). Supervised bilingual lexicon induction with multiple monolingual signals. In *NAACL*.

Justeson J. and Katz S. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Computational Linguistics* **22**(1), 1–38, MIT Press.

Kementchedjhieva Y., Hartmann M. and Søgaard A. (2019). Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *EMNLP*, pp. 3327–3332.

Kilgarriff A., Rychly P., Smrz P. and Tugwell D. (2004). The sketch engine. *Information Technology*, 105–116. See https://www.researchgate.net/profile/Adam_Kilgarriff/publication/260387608_ITRI-04-08_the_sketch_engine/links/54e0d1210cf24d184b0de48f/ITRI-04-08-the-sketch-engine.pdf

Koehn P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, vol. 5, pp. 79–86.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A. and Herbst E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL*, pp. 177–180.

Lin Y., Liu Z., Sun M., Liu Y. and Zhu X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *AAAI*.

Mikolov T., Grave E., Bojanowski P., Puhrsch C. and Joulin A. (2017). Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405.

Mikolov T., Le Q.V. and Sutskever I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Miller G. (1998). *WordNet: An Electronic Lexical Database*, MIT Press.

Nickel M., Rosasco L. and Poggio T. (2019). Holographic embeddings of knowledge graphs. In *AAAI*.

Nguyen D.Q. (2017). An overview of embedding models of entities and relationships for knowledge base completion. arXiv preprint arXiv:1703.08098.

Nguyen D.Q., Nguyen T.D., Nguyen D.Q. and Phung D. (2017). Novel embedding model for knowledge base completion based on convolutional neural network. arXiv preprint arXiv:1712.02121.

Pennington J., Socher R. and Manning C. (2014). GloVe: Global vectors for word representation. In *EMNLP*, pp. 1532–1543.

Rapp R. (1995). Identifying word translations in non-parallel texts. arXiv preprint cmp-lg/9505037.

Ruder S., Vulić I. and Søgaard A. (2017). A survey of cross-lingual word embedding models. arXiv preprint arXiv:1706.04902.

Smadja F., McKeown K. and Hatzivassiloglou V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* **22**(1), 1–38, MIT Press.

Smith S., Turban D., Hamblin S. and Hammerla N. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859.

Sun Y., Wang S., Li Y., Feng S., Tian H., Wu H. and Wang H. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*.

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press

Sun Z., Deng Z.-H., Nie J.-Y. and Tang J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197.

Trouillon T., Welbl J., Riedel S., Gaussier É. and Bouchard G. (2016). Complex embeddings for simple link prediction. *International Conference on Machine Learning (ICML).*

Wang A., Singh A., Michael J., Hill F., Levy O. and Bowman S.R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Wang Z., Zhang J., Feng J. and Chen Z. (2014). Knowledge graph embedding by translating on hyperplanes. In *AAAI*.

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser Ł., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M. and Dean J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Yang B., Yih W.-T., He X., Gao J. and Deng L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Yu S.Y., Rokka Chhetri S., Canedo A., Goyal P., Faruque M.A.A. (2019). Pykg2vec: A python library for knowledge graph embedding. arXiv preprint arXiv:1906.04239.

Cite this article: Church KW (2020). Benchmarks and goals. *Natural Language Engineering* **26**, 579–592. https://doi.org/10.1017/S1351324920000418

https://doi.org/10.1017/S1351324920000418 Published online by Cambridge University Press