2018 12th IEEE International Conference on Semantic Computing

Multilingual Semantic Relatedness using lightweight machine translation

Siamak Barzegar*, Brian Davis†, Siegfried Handschuh‡ and Andre Freitas§
*Insight Centre for Data Analytics, National University of Ireland, Galway, Email: siamak.barzegar@insight-centre.org
†Department of Computer Science, Maynooth University, Email: brian.davis@mu.ie
‡Department of Computer Science and Mathematics, University of Passau, Email: siegfried.handschuh@uni-passau.de
§School of Computer Science, University of Manchester, Email: andre.freitas@manchester.ac.uk

***Abstract*—Distributional semantic models are strongly dependent on the size and the quality of the reference corpora, which embeds the commonsense knowledge necessary to build comprehensive models. While high-quality texts containing large-scale commonsense information are present in English, such as Wikipedia, other languages may lack sufficient textual support to build distributional models. This paper proposes using the combination of a lightweight (sloppy) machine translation model and an English Distributional Semantic Model (DSM) to provide higher quality word vectors for languages other than English. Results show that the lightweight MT model introduces significant improvements when compared to language-specific distributional models. Additionally, the lightweight MT outperforms more complex MT methods for the task of word-pair translation.**

I. Introduction

Distributional Semantic Models (DSM) are consolidating themselves as fundamental components for supporting automatic semantic interpretation in different application scenarios in natural language processing. From *question answering systems*, to *semantic search* and *text entailment*, distributional semantic models support a scalable approach for representing the meaning of words, which can automatically capture comprehensive associative commonsense information by analysing word-context patterns in large-scale corpora in an unsupervised or semi-supervised fashion [1], [2], [3].

However, such DSMs are strongly dependent on the size and the quality of the reference corpora, which embeds the commonsense knowledge necessary to build comprehensive models. While high-quality texts containing large-scale commonsense and domain-specific information are present in English, other languages may lack sufficient textual support to build comprehensive distributional models.

This paper proposes the combination of a lightweight machine translation (MT) model and an English DSM as a mechanism to provide knowledge-rich word vectors for languages other than English. While the problem of delivering high-quality sentence MT requires large parallel corpora and resource-intensive ML models, we claim that the MT for accessing distributional word vectors can be achieved with a lightweight model. In the context of this work, a lightweight MT model is a model which accesses the unigram-level source-target probabilities which can be directly computed from the parallel corpora.

This paper aims at addressing the following research questions:

- Can a lightweight MT model over an English DSM provide higher quality word vectors compared to native word vectors?
- How does a lightweight MT model compares with more complex MT models?
- How parallel corpora size influences the quality of the distributional vector?
- Are there DSMs which are more/less robust to the quality of the MT?

Figure 1 depicts a summary of the experimental model aimed by this paper, where the lightweight MT is compared against state-of-the-art MT services for different word similarity/relatedness datasets.

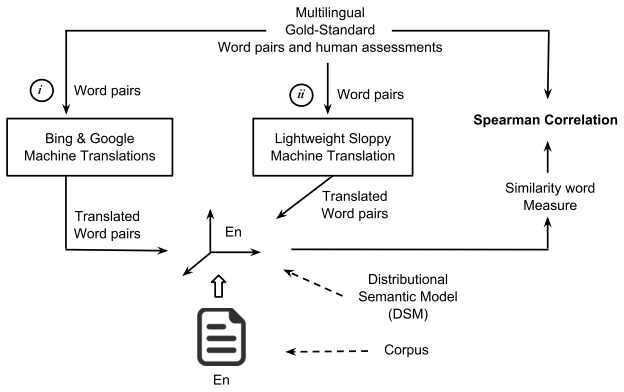

Fig. 1. Depiction of the experimental setup of the experiment.

This paper is organised as follows: Section II describes the related work, Section IV describes the experimental setting, a lightweight machine translation is proposed at section III; while Section V analyses the results and provides the comparative analysis from different models and languages. Finally, Section VI provides the conclusion and future work.

II. Related Work

The majority of related work has concentrated on leveraging joint multi-lingual information to improve the performance of

0-7695-6360-0/18/$31.00 ©2018 IEEE
DOI 10.1109/ICSC.2018.00024

semantic similarity/relatedness models.

Faruqui & Dyer[4] use the distributional invariance across languages and propose a technique based on canonical correlation analysis (CCA) for merging multi-lingual evidence into vectors generated in monolingual fashion. The authors evaluate the resulting word representations on semantic similarity/relatedness evaluation tasks, showing the improvement of multi-lingual over the monolingual scenario.

Utt & Pado[5], develop methods that take advantage of the availability of annotated corpora in English using a translation-based approach to transport the word-link-word co-occurrences to support the creation of syntax-based DSMs. Navigli & Ponzetto[6] propose an approach to compute semantic relatedness exploiting the joint contribution of different languages mediated by lexical and semantic knowledge bases. The proposed model uses a graph-based approach of joint multi-lingual disambiguated senses which outperforms the monolingual scenario and achieves competitive results for both resource-rich and resource-poor languages.

Zou et al.[7] describe an unsupervised semantic embedding (bilingual embedding) for words across two languages that represent semantic information of monolingual words, but also semantic relationships across different languages. The motivation of their work was on the difficulty of identifying semantic similarities across languages, especially when word co-occurrences are rare in the training parallel text. Al-Rfou et al.[8] produced multi-lingual word embeddings for about 100 languages using Wikipedia as the reference corpora.

Freitas et al.[9] investigate how different distributional semantic models built from corpora in different languages and with different sizes perform in computing semantic similarity and relatedness tasks. Additionally, they analysed the role of heavyweight Google and Bing machine translation approaches to support the construction of better distributional vectors and for computing semantic similarity and relatedness measures for other languages. This is the most similar work to our model. Comparatively, this work aims at providing an analysis of the impact of a lightweight machine translation over an English DSM and answering the question on what is the underlying MT quality necessary to deliver word vector models with quality comparable to English.

III. Lightweight Machine Translation

The lightweight MT model is built by processing the set of source—target word alignments within the parallel corpora and by computing the $\omega(s|t)$ word translation table. Given this alignment, it is quite straight-forward to estimate a maximum likelihood lexical translation table.

Given a word pair w_1,w_2 in a language L other than English, the semantic similarity $sim(w_1, w_2)$ will be calculated by first collecting all English translations of w_1 and w_2 into the sets $\mathcal{T}_1$, $\mathcal{T}_2$. For a set which is defined by the cross product of $\mathcal{T}_1$, $\mathcal{T}_2$, the word vectors for each element τ_1^i, τ_2^j are produced ($\vec{\tau_1}^i$, $\vec{\tau_2}^j$). The final similarity score is given by getting the top-most similarity score $sim(\vec{\tau_1}^i, \vec{\tau_2}^j)$.

$$sim(w_1, w_2) = \arg\max_{\tau_1^i, \tau_2^j} sim(\vec{\tau_1}^i, \vec{\tau_2}^j)$$

Algorithm 1 describes the lightweight MT model.

Algorithm 1 The algorithm for computing the semantic similarity between two words with the translation

WP : *word pair (w_1,w_2) in a language other than English*
$\tau_1 \leftarrow$ *Collecting all English translations of* w_1 *from the Lexical translation table.*
$\tau_2 \leftarrow$ *Collecting all English translations of* w_2 *from the Lexical translation table.*
CP : *Cross product of* τ_1 *and* τ_2
for all $pairs \in CP$ **do**:
  $Scores \leftarrow$ *Calculate* $sim(\vec{\tau_1}^i, \vec{\tau_2}^j)$.
end for
Return top-most similarity score in $Scores$

In many cases, users of distributional semantic models need to use the word vectors directly instead of the similarity function (typically the case when using distributional word vectors as features for a machine learning model). An analogous procedure could be used as a disambiguation mechanism when looking up single word vectors. In this case, collocated words in the sentence can serve as a supporting mechanism for disambiguation.

Algorithm 2 shows the variation of the model for looking up distributional vectors for a single word.

Algorithm 2 The algorithm for looking up distributional vectors for a single word as a disambiguation mechanism

$SENT$: *Sentence in a language other than English*
for all $W \in SENT$ **do**:
  MW$\leftarrow$ *Meaningful words in* $SENT$ *related to* W .
  $\tau_w \leftarrow$ *Collecting all English translations of* W *from the Lexical translation table.*
  for all $M \in MW$ **do**:
    $\tau_m \leftarrow$ *Collecting all English translations of* M *from the Lexical translation table.*
  end for
  CP : *Cross product of* τ_w *and* τ_m
  for all $pairs \in CP$ **do**:
    $Scores \leftarrow$ *Calculate* $sim(\vec{\tau_w}^i, \vec{\tau_m}^j)$.
  end for
  $\vec{\tau_w}^i \leftarrow$ Based on the top-most similarity score in $Scores$
end for

IV. Experimental Setup

The experimental setup consists of the instantiation of four distributional semantic models (Explicit Semantic Analysis (ESA) [10], Latent Semantic Analysis (LSA) [11], Word2Vec (W2V) [12] and Global Vectors (GloVe) [13]) in 11 different languages - English, German, French, Italian, Spanish, Portuguese, Dutch, Russian, Swedish, Arabic and Farsi.

The DSMs were generated from Wikipedia dumps (January 2015), which were preprocessed by lowercasing, stemming and removing stopwords. For LSA and ESA, the models were generated using the SSpace Package [14], while W2V and GloVe were generated using the code shared by the respective authors. For the experiment, the vector dimensions for LSA, W2V and GloVe were set to 300 while ESA was defined with 1500 dimensions. The difference of size occurs because ESA is composed of sparse vectors. All models used in the generation process the default parameters defined in each implementation.

Each distributional model was evaluated for the task of computing semantic similarity and relatedness measures using four human-annotated gold standard datasets: Miller & Charles (MC) [15], Rubenstein & Goodenough (RG) [16], WordSimilarity 353 (WS-353) [17] and Simlex-999 [18]. As the four word-pair gold-standards were originally in English, except for some languages available in previous works [19], [20], [18], the word pairs were translated and reviewed with the help of paid professional translators[1], skilled in language data localisation tasks. In the word-pair translation task, in case of word sense ambiguity, the translators were instructed to select the senses which are most related to the other word. In order to support reproducibility and comparability, the datasets are available on the web[2].

As baselines for the lightweight machine translation approach, we used the Google Translate Service and the Microsoft Bing Translation Service. The lightweight MT was generated using three parallel corpora: Europarl, DGT and OpenSubtitle2016 [21]. Table I shows details of the parallel corpora size.

The lightweight MT over DSMs was implemented over the Indra service [22].

TABLE I
DETAILS OF PARALLEL CORPORA SIZE (SCALE OF 10^6).

Parallel Corpora	Parameters	Europarl	DGT	OpenSubtitle2016	All
Source=German Target=English	Sentence Alignments	2	3.2	13.9	19.1
	Source Tokens	45.4	48.4	84.7	178.5
	Target Tokens	53.1	53.1	88.3	194.5
Source=French Target=English	Sentence Alignments	2	3	33.8	38.8
	Source Tokens	53.6	57.7	214.6	325.9
	Target Tokens	51.3	52.8	221.7	325.8
Source=Spanish Target=English	Sentence Alignments	2	3.2	49.9	55.1
	Source Tokens	52.7	60.4	297.4	410.5
	Target Tokens	50.2	52.9	320	423.1
Source=Portuguese Target=English	Sentence Alignments	2	3.2	24.9	30.1
	Source Tokens	51	56.5	147.7	255.2
	Target Tokens	50.3	52.6	160	262.9
Source=Italian Target=English	Sentence Alignments	1.9	3.2	26.3	31.4
	Source Tokens	49	54.6	161.1	264.7
	Target Tokens	50.7	53	172.2	275.9
Source=Swedish Target=English	Sentence Alignments	1.9	3.2	11.9	17
	Source Tokens	42.2	47.1	69.4	158.7
	Target Tokens	46.7	53	81.2	180.9
Source=Dutch Target=English	Sentence Alignments	2	3.2	28.8	34
	Source Tokens	51.2	53.4	182.8	287.4
	Target Tokens	50.6	52.8	197.4	300.8

V. EVALUATION & RESULTS

A. Lightweight Machine Translation vs. Language-Specific Models

In the first part of the experiment we evaluate how the semantic similarity supported by the lightweight MT model performs in comparison to DSMs built over native language corpora. The Spearman Correlation (ρ) between human assessments was calculated for all native-language DSMs and English lookups supported by lightweight MT [9].

The impact of the MT model can be better interpreted by examining the difference between the lightweight machine translation and the language-specific models (depicted in Table III). GLOVE accounts for the largest average percent improvement (78.07%) using the lightweight MT model, while LSA accounts for the lowest value (12.96%). The remaining models accounted for substantial improvements (ESA = 13.84%, W2V = 13.91%).

In terms of improvement per language, Italian achieved the highest percent gains (98.27%), while German accounts for lower results (10.41%). The average improvement for the MT over the language specific model for each word-pair dataset is consistently significant: MC = 23.53%, RG = 16.66%, WS353 = 7.44% and SIMLEX-999 = 71.15%. The results shows in overall the results of lightweight MT outperforms the results of the language-specific models.

Another aspect that we can observe is with regard to which language benefited more from the application of the MT model. The comparative analysis between the models (Table II) indicates that Spanish is the best-performing language (0.59), followed by Swedish (0.57). The lowest Spearman correlation was observed in Dutch (0.50). From the tested DSMs, W2V is consistently the best-performing DSM (0.61).

In terms of impact of the lightweight model for computing the Spearman correlation for different gold-standards: MC, RG and Simlex-999 showed higher percentage improvements when compared to WS-353. The explanation can be found in the fact that the three former datasets focus on similarity computations (thus requiring more sensitive and informative semantic models) while WS-353 targets semantic relatedness.

B. Google and Bing vs. Lightweight Machine Translation based Semantic Relatedness

This section provides a comparative analysis of the lightweight MT model and the Google and Bing Services MT baselines. The Spearman correlation for the lightweight MT approach and their difference in relation to Google & Bing are shown in Table II, IV and V respectively.

In the analysis, word pairs were sent to the baseline machine translation services which translated them to English. The translated words were then used to compute the semantic relatedness using the native English DSMs and their Spearman correlations with the translated pairs were computed.

[1]Global Services for Machine Intelligence, Seehttps://www.lionbridge.com/en-us/global-services-for-machine-intelligence

[2]https://rebrand.ly/multilingual-wordpairs

TABLE II
SPEARMAN CORRELATION FOR THE LIGHTWEIGHT MACHINE TRANSLATION MODELS OVER THE ENGLISH CORPUS.

GS	Model	de	fr	it	nl	pt	sv	es	Model AVG.	GS AVG.
MC	ESA	0.80	0.72	0.70	0.63	0.80	0.72	0.80	0.74	0.76
	LSA	0.72	0.71	0.67	0.65	0.67	0.80	0.78	0.72	
	W2V	0.80	0.86	0.75	0.72	0.82	0.89	0.87	0.82	
	GLOVE	0.79	0.78	0.70	0.61	0.80	0.78	0.82	0.75	
RG	ESA	0.71	0.77	0.68	0.68	0.79	0.73	0.81	0.74	0.72
	LSA	0.60	0.60	0.63	0.62	0.66	0.75	0.72	0.66	
	W2V	0.75	0.78	0.70	0.75	0.78	0.78	0.86	0.77	
	GLOVE	0.69	0.75	0.70	0.63	0.78	0.76	0.80	0.73	
WS353	ESA	0.46	0.41	0.39	0.44	0.44	0.42	0.41	0.42	0.47
	LSA	0.52	0.43	0.45	0.47	0.45	0.47	0.45	0.46	
	W2V	0.66	0.59	0.58	0.61	0.59	0.59	0.60	0.60	
	GLOVE	0.45	0.39	0.37	0.41	0.42	0.41	0.42	0.41	
SIMLEX	ESA	0.21	0.16	0.22	0.19	0.23	0.23	0.24	0.21	0.22
	LSA	0.20	0.16	0.20	0.18	0.21	0.23	0.23	0.20	
	W2V	0.21	0.20	0.23	0.22	0.24	0.27	0.27	0.24	
	GLOVE	0.25	0.20	0.26	0.23	0.27	0.27	0.29	0.25	
	Lang AVG.	0.55	0.53	0.51	0.50	0.56	0.57	0.59	0.54	

TABLE III
DIFFERENCE (%) BETWEEN THE LIGHTWEIGHT MACHINE TRANSLATION MODEL AND THE LANGUAGE-SPECIFIC.

GS	Model	de	fr	it	nl	pt	sv	es	Model AVG.	GS AVG.
MC	ESA	19.35	33.02	29.18	7.98	3.34	34.04	23.45	21.48	26.90
	LSA	2.78	28.21	-3.27	10.03	33.76	15.04	9.49	13.72	
	W2V	15.70	59.27	14.92	17.32	78.78	55.52	10.72	36.03	
	GLOVE	23.27	21.41	82.90	57.73	15.72	37.98	15.53	36.36	
RG	ESA	5.01	31.75	8.79	6.35	25.71	15.62	17.52	15.82	16.73
	LSA	-7.72	26.71	3.03	6.93	51.64	16.14	40.81	19.65	
	W2V	-3.79	18.15	-0.48	10.91	42.95	13.70	17.46	14.13	
	GLOVE	1.07	21.74	21.24	15.24	26.02	23.70	12.39	17.34	
WS353	ESA	8.06	12.65	-1.78	-19.20	0.80	-11.18	7.06	-0.51	6.58
	LSA	7.82	6.24	17.73	-5.02	14.40	6.61	21.27	9.87	
	W2V	15.96	10.05	10.70	2.32	10.07	20.19	11.34	11.52	
	GLOVE	3.71	2.61	-6.86	-4.47	9.68	6.53	26.94	5.45	
SIMLEX	ESA	30.30	-14.49	43.77	25.28	18.70	17.78	8.57	18.56	10.26
	LSA	25.98	-28.37	47.90	-6.55	4.97	21.53	-5.14	8.62	
	W2V	-9.17	-26.35	-0.58	-4.82	-0.84	7.44	-7.86	-6.03	
	GLOVE	28.13	-15.10	37.04	31.12	21.10	32.99	3.85	19.88	
	Lang AVG.	10.41	11.72	19.01	9.45	22.30	19.60	13.34	15.12	

The lightweight MT on average performs equivalently or better than Google and Bing MT (with the exception of WS353 for Google): Google (MC = 6.08%, RG = 0.62%, WS353 = -1.53% and SIMLEX-999 = 2.93%), Bing(MC = 27.75%, RG = 13.38%, WS353 = 5.45% and SIMLEX-999 = 2.65%). A possible explanation for this observed behaviour is that the baselines are MT models supported by language models which target the translation of sentences instead of word pairs.

On average the results show that using lightweight MT is equivalent or slightly better to more sophisticated services. However, there were significant individual variations across languages and the baseline MT services. Portuguese and German achieved the highest percent gains (12.88% and 9.65%, respectively), Google MT outperformed the lightweight MT for French, Dutch and Italian (-8.87%, -7.66% and -2.93%, respectively). But compared with the Bing MT, Italian and German achieved the highest percentage gains (31.72% and 29.04%, respectively), while Bing MT outperforms the lightweight MT for French and Dutch (-6.78% and -4.70%, respectively).

TABLE IV
DIFFERENCE (%) BETWEEN THE LIGHTWEIGHT MACHINE TRANSLATION MODEL AND THE GOOGLE MACHINE TRANSLATION SERVICE.

GS	Model	de	fr	it	nl	pt	sv	es	Model AVG.	GS AVG.
MC	ESA	40.85	4.43	-1.45	-7.64	43.10	7.06	-2.31	12.01	6.08
	LSA	5.79	-2.18	-7.26	-5.98	-6.55	8.58	-2.78	-1.48	
	W2V	10.58	5.11	-7.83	1.69	1.11	8.04	-0.60	2.59	
	GLOVE	39.72	-0.57	-4.91	-6.56	33.49	20.25	-2.91	11.22	
RG	ESA	16.87	2.05	-8.06	-11.24	16.12	9.17	7.54	4.63	0.62
	LSA	-9.88	-6.85	-14.51	-11.08	-1.21	11.41	11.21	-2.99	
	W2V	3.95	-4.81	-13.18	-2.51	-0.69	2.99	14.89	0.09	
	GLOVE	6.78	-4.69	-11.56	-17.35	12.42	11.77	7.89	0.75	
WS353	ESA	2.55	-15.69	-12.00	-4.63	41.20	-5.20	-3.48	0.39	-1.53
	LSA	3.70	-16.90	-8.72	-10.11	16.04	-7.16	1.72	-3.06	
	W2V	4.06	-5.39	-3.42	-7.07	6.12	-3.90	3.80	-0.83	
	GLOVE	3.59	-18.42	-16.70	-10.77	29.65	-5.76	0.10	-2.61	
SIMLEX	ESA	10.38	-21.15	35.73	-5.00	5.34	3.30	25.74	7.76	2.93
	LSA	8.56	-20.09	12.54	-9.17	4.96	0.51	20.30	2.51	
	W2V	4.52	-16.82	3.36	-7.40	0.31	5.02	16.02	0.72	
	GLOVE	2.45	-19.97	11.06	-7.71	4.72	-2.90	17.52	0.74	
	Lang AVG.	9.65	-8.87	-2.93	-7.66	12.88	3.95	7.17	2.03	

TABLE V
DIFFERENCE (%) BETWEEN THE LIGHTWEIGHT MACHINE TRANSLATION MODEL AND THE BING MACHINE TRANSLATION SERVICE.

GS	Model	de	fr	it	nl	pt	sv	es	Model AVG.	GS AVG.
MC	ESA	73.62	6.91	82.99	-12.92	68.40	3.49	25.83	35.47	22.75
	LSA	20.90	3.39	104.06	-12.21	4.45	10.89	22.03	21.93	
	W2V	45.04	13.66	63.67	6.42	16.84	12.61	15.27	24.79	
	GLOVE	65.57	5.81	55.90	-8.95	58.01	9.84	15.37	28.79	
RG	ESA	33.29	3.61	29.67	-5.56	35.02	18.10	17.07	18.74	13.38
	LSA	6.27	1.58	31.33	-5.83	11.02	19.31	6.69	10.05	
	W2V	20.05	-1.14	25.17	7.63	9.38	11.20	8.71	11.57	
	GLOVE	23.92	-2.04	20.84	-8.58	30.83	19.57	7.47	13.14	
WS353	ESA	23.22	-8.38	6.09	0.09	19.99	1.57	-1.28	5.90	5.45
	LSA	26.40	-8.92	13.83	-5.05	14.80	1.94	5.69	6.96	
	W2V	14.31	-5.67	5.89	-1.37	4.88	-1.55	8.58	3.58	
	GLOVE	29.45	-6.35	3.27	-3.79	14.25	-0.17	0.84	5.36	
SIMLEX	ESA	14.74	-31.42	30.73	-12.37	2.34	-3.83	18.25	2.63	2.65
	LSA	22.53	-32.17	5.93	-3.67	0.90	-5.46	15.10	0.45	
	W2V	25.55	-20.04	8.06	-3.31	-0.40	-4.06	20.74	3.79	
	GLOVE	19.84	-27.34	20.14	-5.77	5.76	-4.13	17.66	3.74	
	Lang AVG.	29.04	-6.78	31.72	-4.70	18.53	5.58	12.75	12.31	

C. Word-pair Machine Translation Quality

In order to verify the hypothesis that the translation accuracy of the lightweight model is equivalent or superior to the baseline MT models, the quality of the MT was evaluated in isolation. Tables VI, VII and VIII show the accuracy of all MT approaches using the translated gold-standard. The accuracy of the translation of the lightweight MT significantly outperforms the Bing and Google MT, except for 3 languages, especially for German (-7.89%).

D. Parallel Corpora Size & MT Quality

Our last analysis focuses on the correlation between the size of the supporting parallel corpora used to built the lightweight MT model and the Spearman correlation for each gold standard, averaged for all models (Figure 2). As the lightweight MT model works over a word-based lexical table, the model is more dependent on a parallel corpora with a representative set of unigram translations instead of a language model which is able to model phrasal (above bigram) translations. This shows that the lightweight MT can be potentially transported to languages with smaller parallel corpora.

TABLE VI
TRANSLATION ACCURACY FOR THE LIGHTWEIGHT MT.

dataset/lang	de	fr	it	nl	pt	sv	es	GS AVG.
MC	0.54	0.60	0.61	0.68	0.54	0.66	0.65	0.61
RG	0.48	0.61	0.56	0.61	0.47	0.68	0.64	0.58
WS353	0.83	0.82	0.75	0.86	0.82	0.83	0.80	0.82
SIMLEX	0.74	0.71	0.75	0.78	0.75	0.74	0.76	0.75
Lang AVG	0.65	0.68	0.67	0.73	0.65	0.73	0.71	0.69

TABLE VII
DIFFERENCE (%) IN TRANSLATION ACCURACY BETWEEN LIGHTWEIGHT MT AND GOOGLE MT.

dataset/lang	de	fr	it	nl	pt	sv	es	GS AVG.
MC	-10.77	-2.10	6.12	-12.20	7.69	3.14	1.91	-0.88
RG	-12.35	7.59	3.75	6.33	12.65	3.70	3.30	3.57
WS353	-0.94	3.58	7.50	0.90	-7.72	4.43	-2.13	0.80
SIMLEX	-7.50	4.87	-0.45	-1.45	-6.06	9.01	-4.93	-0.93
Lang AVG	-7.89	3.49	4.23	-1.60	1.64	5.07	-0.46	0.64

TABLE VIII
DIFFERENCE (%) IN TRANSLATION ACCURACY BETWEEN LIGHTWEIGHT MT AND BING MT.

dataset/lang	de	fr	it	nl	pt	sv	es	GS AVG.
MC	12.07	27.68	47.00	17.14	-9.72	10.42	9.03	16.23
RG	8.19	-7.06	38.21	1.28	-7.20	9.69	8.12	7.32
WS353	3.08	-4.57	-2.74	0.58	0.09	0.17	0.18	-0.46
SIMLEX	2.87	-7.98	-0.15	-0.33	-4.30	-2.94	4.89	-1.13
Lang AVG	6.55	2.02	20.58	4.67	-5.28	4.33	5.55	5.49

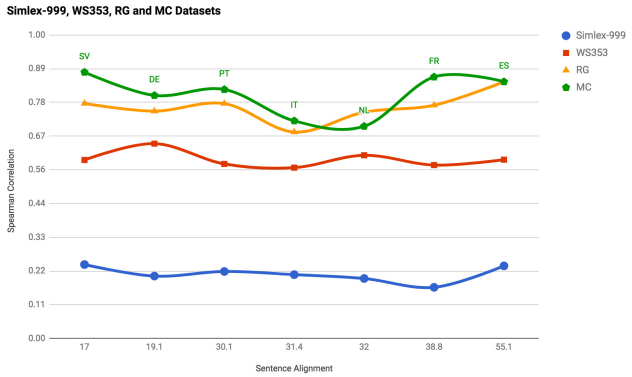

Fig. 2. Correlation between the Spearman correlation values evaluated by lightweight MT over English-DSM and size of parallel corpora that the sloppy MT is learned over them.

VI. CONCLUSION

This paper proposed the use of a lightweight Machine Translation (MT) model over an English Distributional Semantic Model (DSM) as an intermediate layer for the creation of high-quality multi-lingual distributional word vectors. The results show that the proposed model consistently outperforms native language DSMs for word pair similarity evaluation settings: MC (39.12%) , RG (39.59%), WS-353 (14.22%) and SIMLEX-999 (113.41%). Additionally, the paper shows that the lightweight MT model is in the worst case equivalent and in some cases outperforms state-of-the-art MT systems for the translation of word pairs.

Future work will concentrate on the analysis of the suitability of lightweight MT approaches for computing compositional-distributional over phrasal elements.

ACKNOWLEDGMENTS

This publication has emanated from research funded in part from the European Unions Horizon 2020 research and innovation programme under grant agreement No 645425 SSIX and Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

We would like in particular to thank Alexandros Poulis and Juha Vilhunen from the Global Services for Machine Intelligence Group, Lionbridge Finland [3] ensuring the production word of high quality translations for our similarity datasets.

REFERENCES

[1] A. Freitas, "Schema-agnositc queries over large-schema databases: a distributional semantics approach." Ph.D. dissertation, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, 2015.

[2] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Int. Res.*, vol. 37, no. 1, pp. 141–188, Jan. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1861751.1861756

[3] J. E. Sales, A. Freitas, B. Davis, and S. Handschuh, "A compositional-distributional semantic model for searching complex entity categories," in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*, 2016, pp. 199–208.

[4] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 462–471. [Online]. Available: http://www.aclweb.org/anthology/E14-1049

[3] https://www.lionbridge.com/en-us/global-services-for-machine-intelligence

[5] J. Utt and S. Pad, "Crosslingual and multilingual construction of syntax-based vector space models," *Transactions of the Association of Computational Linguistics*, vol. 2, pp. 245–258, 2014.
[6] R. Navigli and S. P. Ponzetto, "Babelrelate! a joint multilingual approach to computing semantic relatedness," in *AAAI Conference on Artificial Intelligence*, 2012.
[7] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation." in *EMNLP*, 2013, pp. 1393–1398.
[8] R. Al-Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual nlp," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 183–192. [Online]. Available: http://www.aclweb.org/anthology/W13-3520
[9] A. Freitas, S. Barzegar, J. E. Sales, S. Handschuh, and B. Davis, "Semantic relatedness for all (languages): A comparative analysis of multilingual semantic relatedness using machine translation," in *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*. Springer, 2016, pp. 212–222.
[10] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611. [Online]. Available: http://dl.acm.org/citation.cfm?id=1625275.1625535
[11] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop Papers*, 2013.
[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, vol. 12, pp. 1532–1543, 2014.
[14] D. Jurgens and K. Stevens, "The s-space package: An open source package for word space models," in *Proceedings of the ACL 2010 System Demonstrations*, ser. ACLDemos '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 30–35. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858933.1858939
[15] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.
[16] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.
[17] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.
[18] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2016.
[19] M. Faruqui and C. Dyer, "Community evaluation and exchange of word vectors at wordvectors. org," 2014.
[20] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "A framework for the construction of monolingual and cross-lingual word similarity datasets," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. Citeseer, 2015, pp. 1–7.
[21] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *LREC*, vol. 2012, 2012, pp. 2214–2218.
[22] S. Barzegar, J. E. Sales, A. Freitas, S. Handschuh, and B. Davis, "Dinfra: A one stop shop for computing multilingual semantic relatedness," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 1027–1028. [Online]. Available: http://doi.acm.org/10.1145/2766462.2767870