Multi-Simlex : une ressource multilingue pour l'évaluation de la notion de similarité sémantique

Thierry Poibeau

LATTICE & Institut 3IA PRAIRIE

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS | PSL

cnrs

Multi-SimLex

- Ressource élaborée principalement à Cambridge en 2019
- Publication (Arxiv, mars 2020, puis *Computational Linguistics*)

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart and Anna Korhonen. 2021. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity. *Computational Linguistics* (2021) 46 (4): 847–897. https://doi.org/10.1162/coli_a_00391

Similarité sémantique

- Similarité <> Association / relation sémantique <> synonymie
  - Voiture ~ automobile < fourgonnette < camion
  - Voiture : roue volant rouler décapotable

- Similarité sémantique : notion fondamentale en TAL (et en linguistique)
- Les modèles de langage (ex. Word2Vec) repèrent des associations sémantique, mais quid de la similarité ?

Les modèles distinguent-ils la similarité des autres relations sémantiques ?

- Evaluer la capacité des modèles de langage à repérer des similarités sémantiques
  - Similarité vs non similarité
  - Proximité sémantique (notion de distance entre mots)

- Nécessité de ressources pour mesurer la similarité sémantique
  - En fonction de la langue
  - En fonction de la fréquence des mots pris en compte
  - En fonction des parties du discours (nom / verbe/ adjectif / adverbe)

Ressources antérieures : SimLex et SimVerb

- SimLex-999 (Hill, Reichart, and Korhonen 2015)
  - 999 couples de mots
  - Essentiellement des noms
  - Principalement des mots très fréquents
  - Ressource très populaire (+ de 1000 citations dans Google Scholar)
- SimVerb-3500 (Gerz et al. 2016)
  - Idem pour les verbes
  - Contenu aspectuel des verbes

Autre ressources

- Peu de choses sur la similarité
- Dictionnaires de synonymes (sans score)
- WordSim (Finkelstein et al., 2007) : relations sémantiques divers entre noms (lien sémantique) <> SimLex

Pair	**Simlex-999 rating**	**WordSim-353 rating**
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

Limites de Simlex et SimVerb

- Ressources disponibles essentiellement pour l'anglais
  - Quelques versions dans d'autres langues, mais non directement comparables (traduction d'une partie de la base seulement, liste de couples de mots différentes, etc)
- Essentiellement des mots très fréquents
- Essentiellement des noms et des verbes

➢ Multi-SimLex

Multi-Simlex : aperçu

- Disponible en 12 langues
  - Langues indo-européennes et non indo-européennes
  - Langues « dominantes » et langues sous-dotées
- Contenu lexical similaire d'une langue à l'autre
- Normalisation de la mise au point des ressources pour les différentes langues cibles
  - Protocoles et guides d'annotation clairs
- Un protocole ouvert, extensible à d'autres langues
  - Cf. Modèle Universal Dependencies

Critères et principes préalables

- (C1) Représentativité et diversité. Différentes partie du discours, différents types de concepts (concrets et abstraits), de domaines et de fréquences différentes.
- (C2) Contenu clairement défini. La relation à annoter doit être clairement définie, éventuellement en contraste avec d'autres relations.
- (C3) Consistance et fiabilité. Les annotations doivent pouvoir provenir de locuteurs non experts à partir de guide d'annotation fiables et précis. L'accord entre annotateurs doit être mesurable.

Elaboration de la ressource

- Etablissement du référentiel pour l’anglais
  - Liste de couples de mots, échelle de notation des liens
  - Protocoles et guides d’annotation
- Choix des langues cible
- Pour chaque langue cible
  - Traduction de la base
  - Affectation des scores et adjudication

Les langues visées

Language	ISO 639-3	Family	Type	# Speakers
Chinese Mandarin	CMN	Sino-Tibetan	Isolating	1.116 B
Welsh	CYM	IE: Celtic	Fusional	0.7 M
English	ENG	IE: Germanic	Fusional	1.132 B
Estonian	EST	Uralic	Agglutinative	1.1 M
Finnish	FIN	Uralic	Agglutinative	5.4 M
French	FRA	IE: Romance	Fusional	280 M
Hebrew	HEB	Afro-Asiatic	Introflexive	9 M
Polish	POL	IE: Slavic	Fusional	50 M
Russian	RUS	IE: Slavic	Fusional	260 M
Spanish	SPA	IE: Romance	Fusional	534.3 M
Kiswahili	SWA	Niger-Congo	Agglutinative	98 M
Yue Chinese	YUE	Sino-Tibetan	Isolating	73.5 M

Table 1: The list of 12 languages in the Multi-SimLex multilingual suite along with their corresponding language family (IE = Indo-European), broad morphological type, and their ISO 639-3 code. The number of speakers is based on the total count of L1 and L2 speakers, according to ethnologue.com.

Traduction vers une langue cible

- Toute traduction doit être unique (pas de duplication de couple de mots)
- Les deux mots d'un couple doivent être différents (on ne peut pas traduire *car* et *automobile* en espagnol par *coche*).
- La traduction qui rend le mieux la relation de similarité entre les deux mots à traduire doit être privilégiée
- S'il n'y a pas de mot simple équivalent dans la langue cible, il est possible de recourir à un mot composé (*praca domowa* pour *homework* en polonais).

Problèmes de traduction

- Deux traductions faites indépendamment + phase d'adjudication
- Problèmes typiques
  - Mots polysémiques (verbes)
  - Distinction propres à l'anglais
  - Particularités linguistiques : nom de parenté en chinois (*petit frère*, *grand frère* pour *frère*), *Wood* vs *timber* en estonien *puit*
  - Cf. détails pour le français à la suite

Accord inter-annotateur

Languages:	CMN	CYM	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE	*Avg*
Nouns	84.5	80.0	90.0	87.3	78.2	98.2	90.0	95.5	85.5	80.0	77.3	86.0
Adjectives	88.5	88.5	61.5	73.1	69.2	100.0	84.6	100.0	69.2	88.5	84.6	82.5
Verbs	88.0	74.0	82.0	76.0	78.0	100.0	74.0	100.0	74.0	76.0	86.0	82.5
Adverbs	92.9	100.0	57.1	78.6	92.9	100.0	85.7	100.0	85.7	85.7	78.6	87.0
Overall	86.5	81.0	82.0	82.0	78.0	99.0	85.0	97.5	80.5	81.0	80.5	84.8

Table 2: Inter-translator agreement (% of matched translated words) by independent translators using a randomly selected 100-pair English sample from the Multi-SimLex dataset, and the corresponding 100-pair samples from the other datasets.

- Scores assez homogènes
- Résultats très élevés pour certaines langues : variation dans la procédure utilisée ? (cf. hébreu, russe)

Bilan sur la traduction

- Beaucoup de couples « faciles » à traduire, avec un bon accord entre annotateurs
- Mais aussi une proportion non négligeable de couples difficile à traduire
  - Pas d'équivalent direct ou de couple similaire (*soccer – football, meter – yard, taxi – cab*)
  - Mots très polysémique (*get,* cf. *get – remain*)
  - Mots difficiles à traduire hors contexte (*log ⇒ rondin; weather ⇒ temps, météo ?*)
  - Cas des verbes (*get – remain; argue – persuade*)
  - Tentation de périphrase pour les adverbes (*d'une manière X*)
  - Enigmes / collocations (*princess – biscuit ⇒ prince – biscuit*)
  - Couple morphologique (*currency – concurrency*)

Affectation des scores de similarité

1. Chaque couple de mot reçoit un score compris entre 0 et 6. 6 = très grande similarité (synonymie parfaite), 0 = similarité nulle.
2. Chaque annotateur doit noter l'ensemble des 1 888 paires de l'ensemble de données (en 2-3 semaines max).
3. Les annotateurs peuvent utiliser des sources externes (dictionnaires, des thésaurus, WordNet) si nécessaire seulement.
4. Les annotateurs ne doivent pas communiquer entre eux pendant le processus d'annotation.
5. Les annotateurs sont rémunérés pour ce travail.

Protocole utilisé

- Un « coordinateur » par langue
- 3 étapes
  - Annotation par chaque annotateur
  - Retour sur les scores les plus différents de la moyenne (par le coordinateur) -> possibilité de correction (pour supprimer les erreurs manifestes)
  - Si plus de 10 annotateurs, possibilité de supprimer les annotateurs avec le score le plus bas comparé à la moyenne des autres annotateurs

Annotateurs par langue et accord entre annotateurs

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
R1: Start	13	12	14	12	13	10	11	12	12	12	11	13
R3: End	11	10	13	10	10	10	10	10	10	10	10	11

Table 3: Number of human annotators. R1 = Annotation Round 1, R3 = Round 3.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Nouns	0.661	0.622	0.659	0.558	0.647	0.698	0.538	0.606	0.524	0.582	0.626	0.727
Adjectives	0.757	0.698	0.823	0.695	0.721	0.741	0.683	0.699	0.625	0.64	0.658	0.785
Verbs	0.694	0.604	0.707	0.58	0.644	0.691	0.615	0.593	0.555	0.588	0.631	0.76
Adverbs	0.699	0.593	0.695	0.579	0.646	0.595	0.561	0.543	0.535	0.563	0.562	0.716
Overall	0.68	0.619	0.698	0.583	0.646	0.697	0.572	0.609	0.53	0.576	0.623	0.733

Table 4: Average pairwise inter-annotator agreement (APIAA). A score of 0.6 and above indicates strong agreement.

Répartition des couples de mots par intervalle

Lang:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Interval												
$[0, 1)$	56.99	52.01	50.95	35.01	47.83	17.69	28.07	49.36	50.21	43.96	61.39	57.89
$[1, 2)$	8.74	19.54	17.06	30.67	21.35	20.39	35.86	17.32	22.40	22.35	11.86	7.84
$[2, 3)$	13.72	11.97	12.66	16.21	12.02	22.03	16.74	11.86	11.81	14.83	9.11	11.76
$[3, 4)$	11.60	8.32	8.16	10.22	10.17	17.64	8.47	8.95	8.10	9.38	7.10	12.98
$[4, 5)$	6.41	5.83	6.89	6.25	5.61	12.55	6.62	7.57	5.88	6.78	6.30	6.89
$[5, 6]$	2.54	2.33	4.29	1.64	2.97	9.64	4.24	4.93	1.59	2.70	4.24	2.65

Table 6: Fine-grained distribution of concept pairs over different rating intervals in each Multi-SimLex language, reported as percentages. The total number of concept pairs in each dataset is 1,888.

Comparaison des scores entre langues

- Rôle pivot de l'anglais

	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA
CYM	0.725										
ENG	0.778	0.827									
EST	0.740	0.771	0.823								
FIN	0.714	0.768	0.800	0.776							
FRA	0.723	0.767	0.820	0.778	0.766						
HEB	0.696	0.737	0.779	0.738	0.736	0.753					
POL	0.718	0.772	0.819	0.792	0.769	0.757	0.730				
RUS	0.696	0.719	0.780	0.763	0.730	0.730	0.731	0.770			
SPA	0.708	0.751	0.801	0.747	0.732	0.756	0.714	0.762	0.733		
SWA	0.627	0.669	0.663	0.645	0.650	0.629	0.633	0.637	0.631	0.633	
YUE	0.861	0.711	0.747	0.717	0.704	0.697	0.686	0.689	0.674	0.688	0.628

Figure 1: Spearman's correlation coefficient (ρ) of the similarity scores for all languages in Multi-SimLex.

Variété des scores par langue

Word Pair	POS	ENG	SPA	SWA	CYM
Similar average rating					
unlikely – friendly	ADV	0	0	0	0
book – literature	N	2.5	2.3	2.1	2.3
vanish – disappear	V	5.2	5.3	5.5	5.3
Different average rating					
regular – average	ADJ	4	4.1	0.5	0.8
care – caution	N	4.1	5.7	0.2	3.1
One language higher					
large – big	ADJ	5.9	2.7	3.8	3.8
bank – seat	N	0	5.1	0	0.1
sunset - evening	N	1.6	1.5	5.5	2.8
purely – completely	ADV	2.3	2.3	1.1	5.4
One language lower					
woman – wife	N	0.9	2.9	4.1	4.8
amazingly – fantastically	ADV	5.1	0.4	4.1	4.1
wonderful – terrific	ADJ	5.3	5.4	0.9	5.7
promise – swear	V	4.8	5.3	4.3	0

Table 7: Examples of concept pairs with their similarity scores from four languages. For brevity, only the original English concept pair is included, but note that the pair is translated to all target languages, see §5.1.

Français

- Improbable – amical : 0
- Livre – littérature 3,2
- Se dissiper – disparaître : 5,5

- Ordinaire – moyen : 3,6
- Soin – attention : 5

- Grand – gros : 3,8
- Siège – banque : 1,9
- Coucher de soleil – soir : 3,8
- Strictement – complètement : 3,2

- Femme – épouse : 4,8
- Incroyablement – fantastiquement : 5,3
- Merveilleux – terrifiant : 1,8
- Promettre – jurer : 5,4

Explications

- Subjectivité de la tâche de scoring
- Non équivalence des concepts / différences entre langue
- Autres phénomènes comme la métonymie, collocations

Jeu de données crosslingues (bilingues)

- Croiser deux jeux de données : *attroupement – foule* (fr) / *rahvasumm – rahvahulk* (est.) ⇒ *attroupement – rahvaluk ; rahvasumm – foule*
- Ne tenir que les couples avec un score différant de moins de 1.2 points

	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
CMN	1,888	–	–	–	–	–	–	–	–	–	–	–
CYM	3,085	1,888	–	–	–	–	–	–	–	–	–	–
ENG	3,151	3,380	1,888	–	–	–	–	–	–	–	–	–
EST	3,188	3,305	3,364	1,888	–	–	–	–	–	–	–	–
FIN	3,137	3,274	3,352	3,386	1,888	–	–	–	–	–	–	–
FRA	2,243	2,301	2,284	2,787	2,682	1,888	–	–	–	–	–	–
HEB	3,056	3,209	3,274	3,358	3,243	2,903	1,888	–	–	–	–	–
POL	3,009	3,175	3,274	3,310	3,294	2,379	3,201	1,888	–	–	–	–
RUS	3,032	3,196	3,222	3,339	3,257	2,219	3,226	3,209	1,888	–	–	–
SPA	3,116	3,205	3,318	3,312	3,256	2,645	3,256	3,250	3,189	1,888	–	–
SWA	2,807	2,926	2,828	2,845	2,900	2,031	2,775	2,819	2,855	2,811	1,888	–
YUE	3,480	3,062	3,099	3,080	3,063	2,313	3,005	2,950	2,966	3,053	2,821	1,888

Table 11: The sizes of all monolingual (main diagonal) and cross-lingual datasets.

Evaluation de modèles de langage avec Multi-SimLex

- Modèles statiques
  - FastText
- Modèles dynamiques
  - Bert
  - XLM
- Version monolingue et multilingue (M-Bert)
- Version sans post-traitement et avec divers post-traitements (voir article)

Conclusion sur l’évaluation

- Multi-SimLex répond à un besoin et permet une évaluation fine des modèles de langage
- Permet de valider certaines hypothèses
  - Plus il y a de données d’entraînement, mieux c’est
  - Les modèles multilingues ne sont pas très performants pour le calcul de la similarité, même avec post-traitements
  - etc.
- Mais supplanté par de nouveaux paradigmes d’évaluation
  - Glue, SuperGlue
  - Jeu de données dynamiques, avec retour des humains dans la course (https://dynabench.org/)

Conclusion générale

- Multi-SimLex est une ressource intéressante pour évaluer la notion de similarité sémantique
- Mais, évolution des façons d'évaluer
- Subjectivité de la tâche
  - Couples de mots hors contexte
  - Traduction parfois sujette à discussion
  - Idem pour l'affectation des scores

- Merci de votre attention