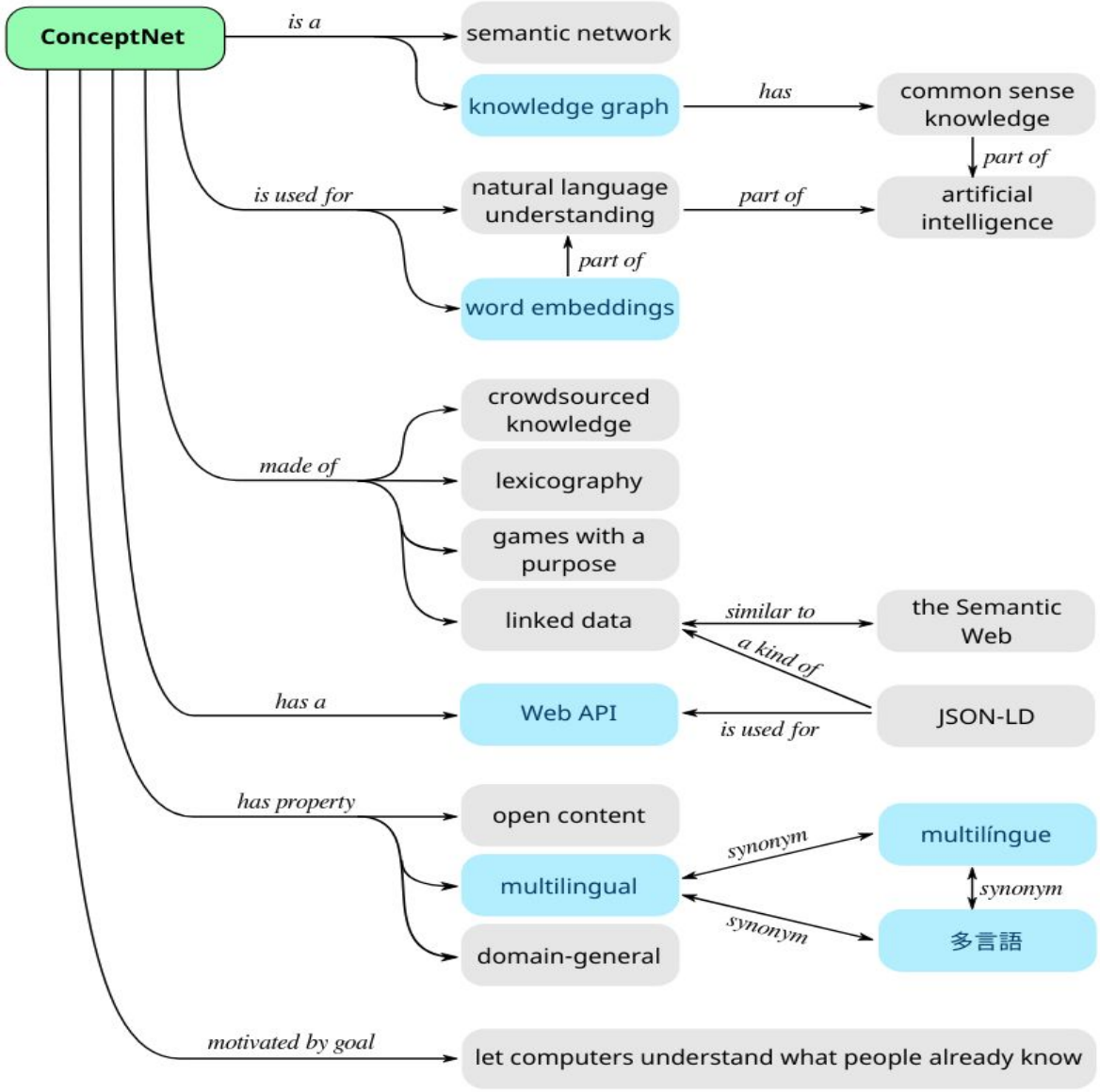
ConceptNet 5.5: An Open Multilingual Graph of General Knowledge

Marcelo Akira Inuzuka
Instituto de Informática - UFG
Professor

CONCEPTNET:

VISÃO GERAL



Para saber mais:
http://conceptnet.io

[Submitted on 12 Dec 2016 (v1), last revised 11 Dec 2018 (this version, v2)]

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge

Robyn Speer, Joshua Chin, Catherine Havasi

Machine learning about language can be improved by supplying it with specific knowledge and sources of external information. We present here a new version of the linked open data resource ConceptNet that is particularly well suited to be used with modern NLP techniques such as word embeddings.
ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose. It is designed to represent the general knowledge involved in understanding language, improving natural language applications by allowing the application to better understand the meanings behind the words people use.
When ConceptNet is combined with word embeddings acquired from distributional semantics (such as word2vec), it provides applications with understanding that they would not acquire from distributional semantics alone, nor from narrower resources such as WordNet or DBPedia. We demonstrate this with state-of-the-art results on intrinsic evaluations of word relatedness that translate into improvements on applications of word vectors, including solving SAT-style analogies.

Para saber mais: https://arxiv.org/abs/1612.03975

→ isA → Knowledge Graph

É uma rede que representa relações semânticas entre conceitos (semantic network)

É um grafo formado por nós e arestas, que respectivamente representam palavras ou frases ligadas a asserções.

É uma base de conhecimento de coisas que as pessoas sabem e os computadores deveriam saber (senso comum)

Para saber mais: https://github.com/commonsense/conceptnet5/wiki/FAQ

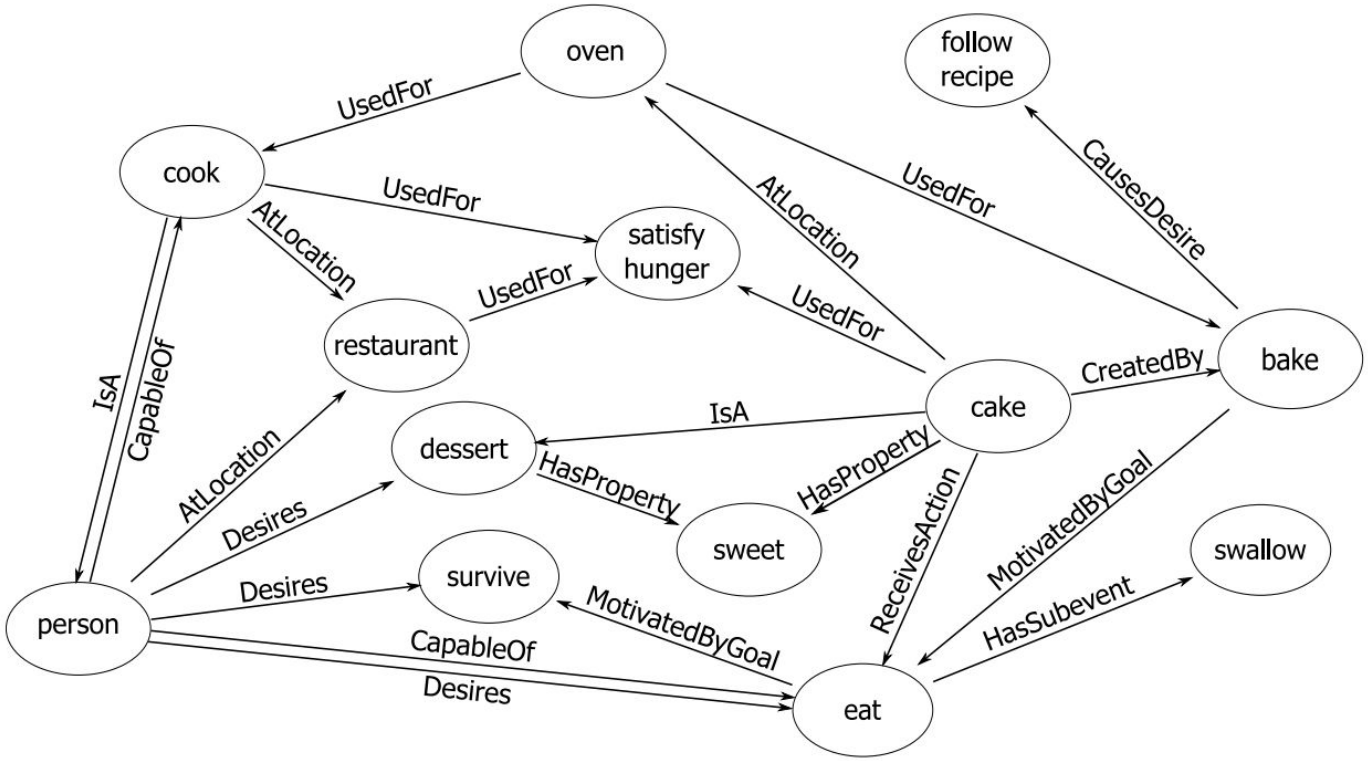


Figure 1: A high-level view of the knowledge ConceptNet has about a cluster of related concepts.

Fonte: http://lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf

Exemplo 1 de 4:

ConceptNet represents relations between words such as:

- A *net* is used for *catching fish*.
- "*Leaves*" is a form of the word "*leaf*".
- The word *cold* in English is *studený* in Czech.
- O *alimento* é usado para *comer* [Food is used for eating].

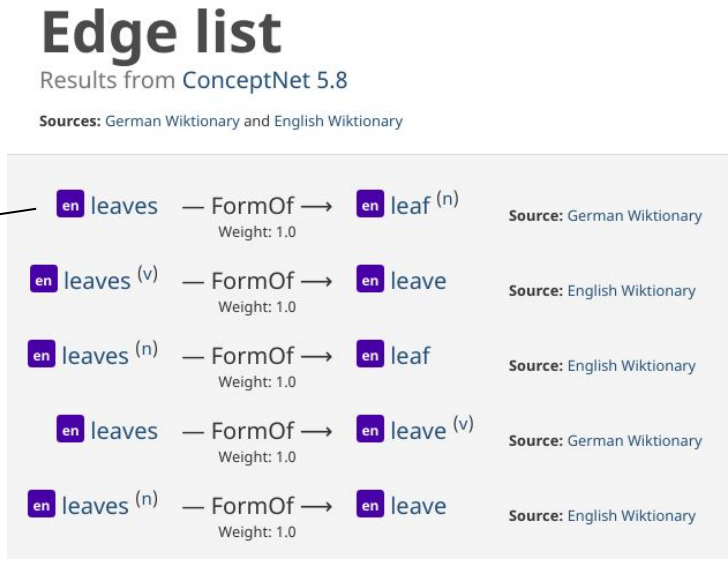
http://conceptnet.io/c/en/net

In this paper, we will concisely represent assertions such as the above as triples of their start node, relation label, and end node: the assertion that "a dog has a tail" can be represented as (*dog*, *HasA*, *tail*).

Exemplo 2 de 4:

ConceptNet represents relations between words such as:

- A *net* is used for *catching fish*.
- "*Leaves*" is a form of the word "*leaf*".
- The word *cold* in English is *studený* in Czech.
- O *alimento* é usado para *comer* [Food is used for eating].

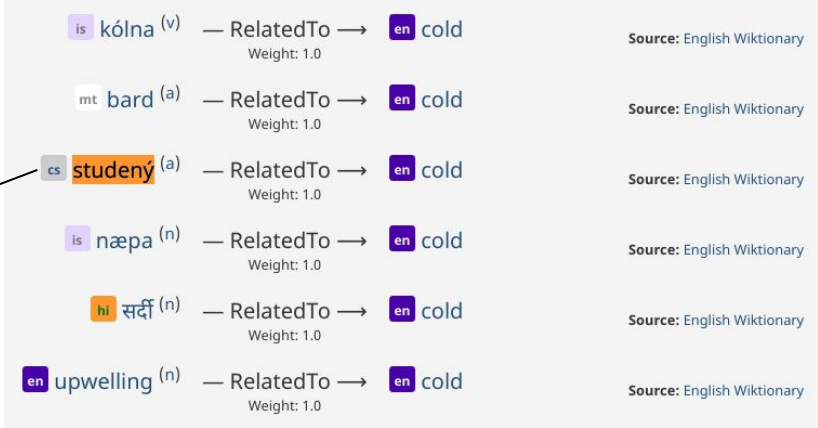


http://conceptnet.io/c/en/leaves?rel=/r/FormOf

Exemplo 3 de 4:

ConceptNet represents relations between words such as:

- A *net* is used for *catching fish*.
- "*Leaves*" is a form of the word "*leaf*".
- The word *cold* in English is *studený* in Czech.
- O *alimento* é usado para *comer* [Food is used for eating].



http://conceptnet.io/c/en/cold?rel=/r/RelatedTo&limit=1000

Exemplo 4 de 4:

ConceptNet represents relations between words such as:

- A *net* is used for *catching fish*.
- "*Leaves*" is a form of the word "*leaf*".
- The word *cold* in English is *studený* in Czech.
- O *alimento* é usado para *comer* [Food is used for eating].

http://conceptnet.io/c/pt/alimento?rel=/r/UsedFor&limit=1000

Fontes de conhecimento

Knowledge Sources

ConceptNet 5.5 is built from the following sources:

- Facts acquired from Open Mind Common Sense (OMCS) (Singh 2002) and sister projects in other languages (Anacleto et al. 2006)
- Information extracted from parsing Wiktionary, in multiple languages, with a custom parser ("Wikiparsec")
- "Games with a purpose" designed to collect common knowledge (von Ahn, Kedia, and Blum 2006) (Nakahara and Yamada 2011) (Kuo et al. 2009)
- Open Multilingual WordNet (Bond and Foster 2013), a linked-data representation of WordNet (Miller et al. 1998) and its parallel projects in multiple languages
- JMDict (Breen 2004), a Japanese-multilingual dictionary
- OpenCyc, a hierarchy of hypernyms provided by Cyc (Lenat and Guha 1989), a system that represents common sense knowledge in predicate logic
- A subset of DBPedia (Auer et al. 2007), a network of facts extracted from Wikipedia infoboxes

Tamanho

With the combination of these sources, ConceptNet contains over 21 million edges and over 8 million nodes. Its English vocabulary contains approximately 1,500,000 nodes, and there are 83 languages in which it contains at least 10,000 nodes.

The largest source of input for ConceptNet is Wiktionary, which provides 18.1 million edges and is mostly responsible for its large multilingual vocabulary. However, much of the character of ConceptNet comes from OMCS and the various games with a purpose, which express many different kinds of relations between terms, such as *PartOf* ("a wheel is part of a car") and *UsedFor* ("a car is used for driving").

Relações

Relations

ConceptNet uses a closed class of selected relations such as *IsA*, *UsedFor*, and *CapableOf*, intended to represent a relationship independently of the language or the source of the terms it connects.

ConceptNet 5.5 aims to align its knowledge resources on its core set of 36 relations. These generalized relations are similar in purpose to WordNet's relations such as *hyponym* and *meronym*, as well as to the qualia of the Generative Lexicon theory (Pustejovsky 1991). ConceptNet's edges are directed, but as a new feature in ConceptNet 5.5, some relations are designated as being symmetric, such as *SimilarTo*. The directionality of these edges is unimportant.

Core relations

The core relations are:

- **Symmetric relations**: *Antonym*, *DistinctFrom*, *EtymologicallyRelatedTo*, *LocatedNear*, *RelatedTo*, *SimilarTo*, and *Synonym*
- **Asymmetric relations**: *AtLocation*, *CapableOf*, *Causes*, *CausesDesire*, *CreatedBy*, *DefinedAs*, *DerivedFrom*, *Desires*, *Entails*, *ExternalURL*, *FormOf*, *HasA*, *HasContext*, *HasFirstSubevent*, *HasLastSubevent*, *HasPrerequisite*, *HasProperty*, *InstanceOf*, *IsA*, *MadeOf*, *MannerOf*, *MotivatedByGoal*, *ObstructedBy*, *PartOf*, *ReceivesAction*, *SenseOf*, *SymbolOf*, and *UsedFor*

Definitions and examples of these relations appear in a page of the ConceptNet 5.5 documentation[1].

Relations with specific semantics, such as *UsedFor* and *HasPrerequisite*, tend to connect common words and phrases, while rarer words are connected by more general relations such as *Synonym* and *RelatedTo*.

Veja mais:
https://github.com/commonsense/conceptnet5/wiki/Relations

Origem e evolução

- Open Mind Common Sense (Singh 2002)
- ConceptNet (Liu and Singh 2004)
- ConceptNet 5 (2012)
- ConceptNet 5.5 (2017)
  - ConceptNet Numberbatch (2017) (word embeddings)
- ConceptNet 5.8 (2020)

Para ficar atualizado: http://blog.conceptnet.io/

Representação de termos

- Uso de unicode => internacionalização
- Tokens separados por '_'. Exemplo: United_States
- Conceitos são pré-fixados com /c/lang, onde lang é o BCP 47 language code. Exemplo: /c/en/United_States
- Relações são pré-fixadas com /r, sem variação de língua. Exemplo: /r/partOf
- Lemmatizer foi retirado da versão 5.4 para 5.5. Exemplo: driving e drive co-existem e são relacionados por uma tupla (driving, formOf, drive)

Desambiguação

- Exemplo, o termo *lead* pode assumir significados diferentes:
  - geral: /c/en/lead
  - como um metal: /c/en/lead/n/wn/substance
  - como um verbo: /c/en/lead/v
  - relacionar-se a vários contextos:
    - http://conceptnet.io/c/en/lead?rel=/r/HasContext&limit=1000
  - vários tipos:
    - http://conceptnet.io/c/en/lead?rel=/r/IsA&limit=1000

Dados ligados

Linked Data

ConceptNet imports knowledge from some other systems, such as WordNet, into its own representation. These other systems have their own target vocabularies that need to be aligned with ConceptNet, which is usually an underspecified, many-to-many alignment.

A term that is imported from another knowledge graph will be connected to ConceptNet nodes via the relation *ExternalURL*, pointing to an absolute URL that represents that term in that external resource. This newly-introduced relation preserves the provenance of the data and enables looking up what the untransformed data was. ConceptNet terms can also be represented as absolute URLs, so this allows ConceptNet to connect bidirectionally to the broader ecosystem of Linked Open Data.

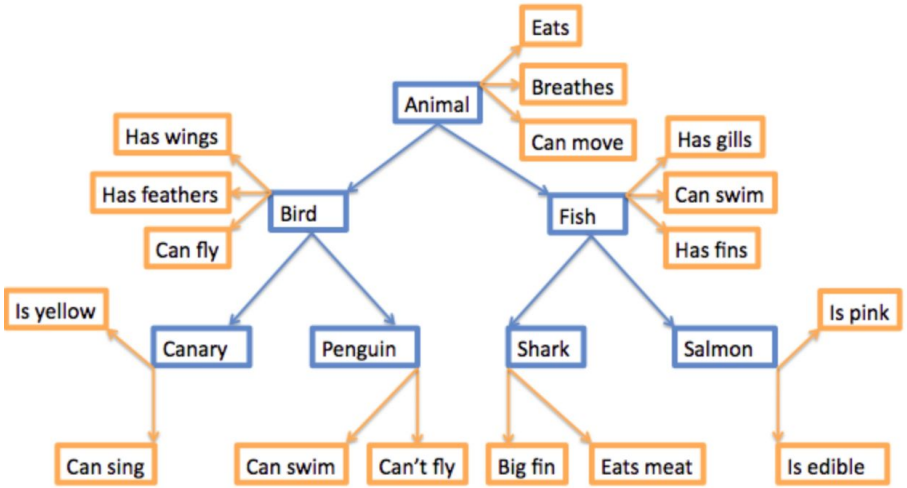
Construindo Word Embeddings

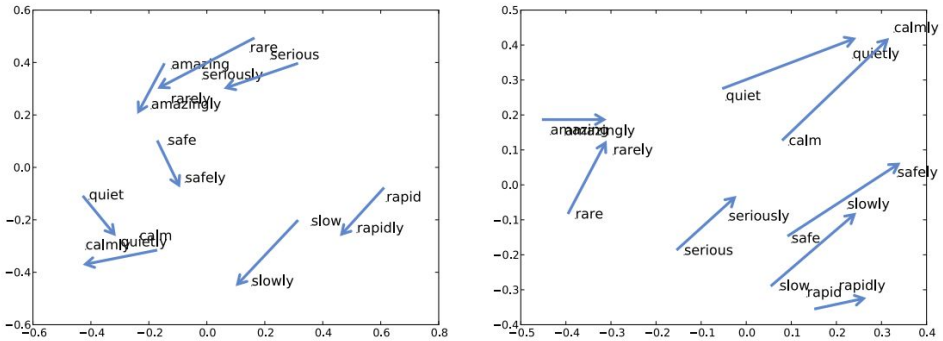
- O grafo do Conceptnet pode ser representado por uma matriz esparsa, simétrica, termo-a-termo:
  - Cada célula contém a soma dos pesos de todas as arestas que conectam dois termos correspondentes
  - Por questão de performance, foram descartados termos com menos de três arestas
  - A matriz representa os termos e seu contexto, análogo a um texto, onde o termo ficaria próximo ao contexto formado por outros termos
  - Para gerar os word embeddings, foi seguido as recomendações de Levy, Goldberg and Dagan (2015)
  - Resultado: ConceptNet-PPMI

Combinando Conceptnet com outros Word Embeddings que aprenderam a partir de texto

- Problema: representações vetoriais de palavras (como um Knowledge Graph) são aprendidas a partir de informações distribucionais de grandes corpora. Apesar de carregar informação semântica, tais representações desprezam informações valiosas de dicionários léxicos-semânticos, tais como WordNet, FrameNet, e o Paraphrase Database.
- Solução: Retrofitting (Faruqui et al. 2015) - processo de refinar Knowledge Graphs com semantic lexicons

Visualização do processo de Retrofitting





Semantic Lexicon

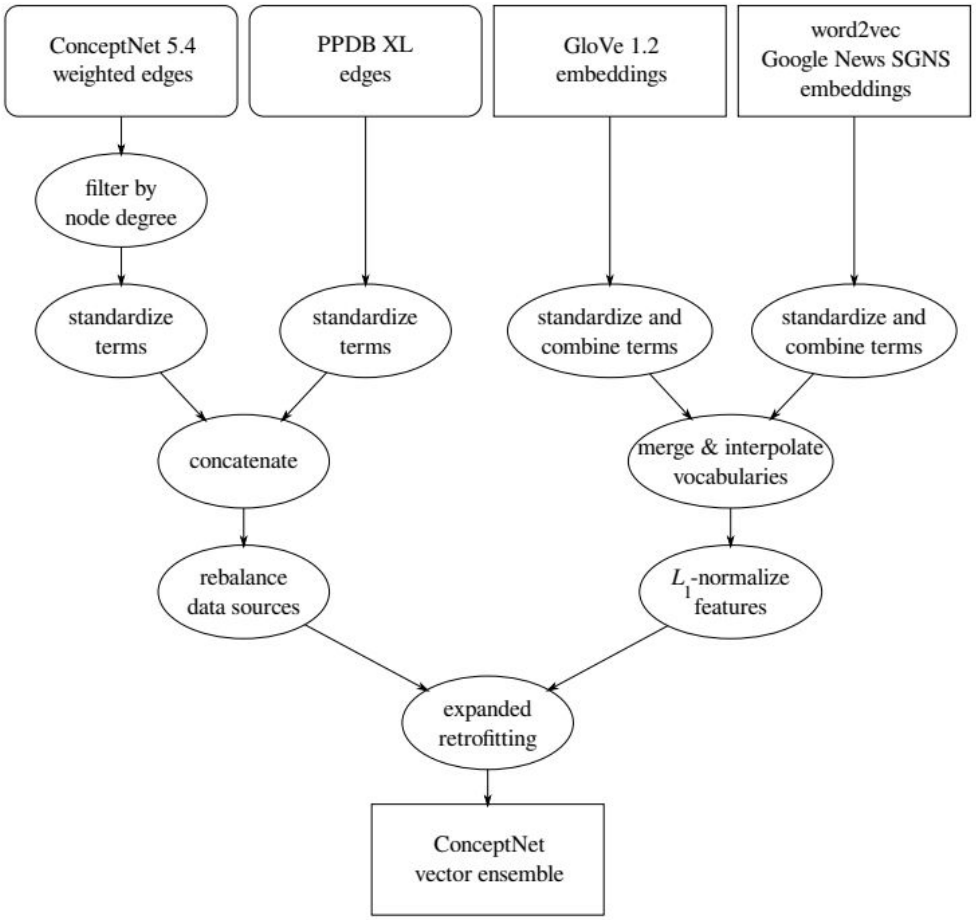
Retrofitting

Figure 3: Two-dimensional PCA projections of 100-dimensional SG vector pairs holding the "adjective to adverb" relation, before (left) and after (right) retrofitting.

Combinando Conceptnet com GloVe, word2vec e PPDB

- Hipótese: combinar Conceptnet com outros distributional word embeddings (GloVe, word2vec e PPDB) que possuem forças complementares, através de um método inspirado em retrofitting, melhora o desempenho?
- Investigação: expanded retrofitting (Speer and Chin 2016)
  - Conceptnet Numberbatch (Conceptnet vector ensemble) = sistema híbrido
  - SOTA em word-similarity / world-relatedness evaluations (café é próximo de caneca, mas café não é um caneca)

Ensemble de embeddings



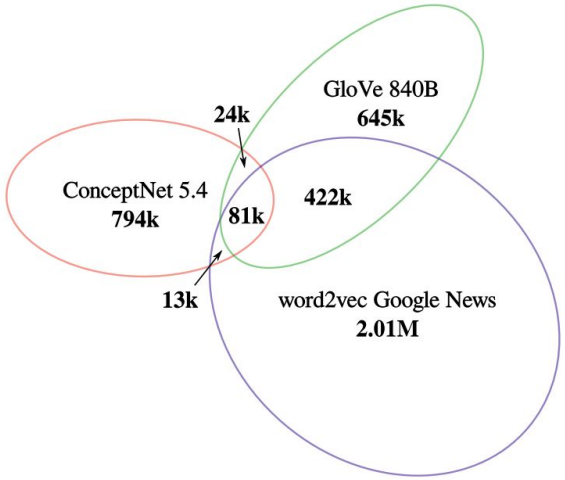


Figure 2: A proportional-area diagram showing the overlap of vocabularies among ConceptNet and the available embeddings for word2vec and GloVe.

Avaliação do ConceptNet Numberbatch

- ConceptNet Numberbatch 16.09 = versão de setembro de 2016
- Comparações com outros word embeddings:

  - word2vec SGNS (Mikolov et al. 2013), trained on Google News text
  - GloVe 1.2 (Pennington, Socher, and Manning 2014), trained on the Common Crawl
  - LexVec (Salle, Idiart, and Villavicencio 2016), trained on the English Wikipedia and NewsCrawl 2014
  - ConceptNet-PPMI, described here and trained on ConceptNet 5.5 alone
  - ConceptNet Numberbatch 16.09, the hybrid of ConceptNet 5.5, word2vec, and GloVe described here

Resultados do ConceptNet Numberbatch

- Word relatedness: MEM-300, Rare Words, MTurk-771 e WS353
- Story Cloze (Mostafazadeh et al. 2016) é uma avaliação de senso comum para avaliar se um método consegue adivinhar o final de uma história escolhendo uma alternativa.
- SAT (ENEM) é um teste de analogia entre pares,por exemplo: “fogo está calor, assim como gelo está para …”

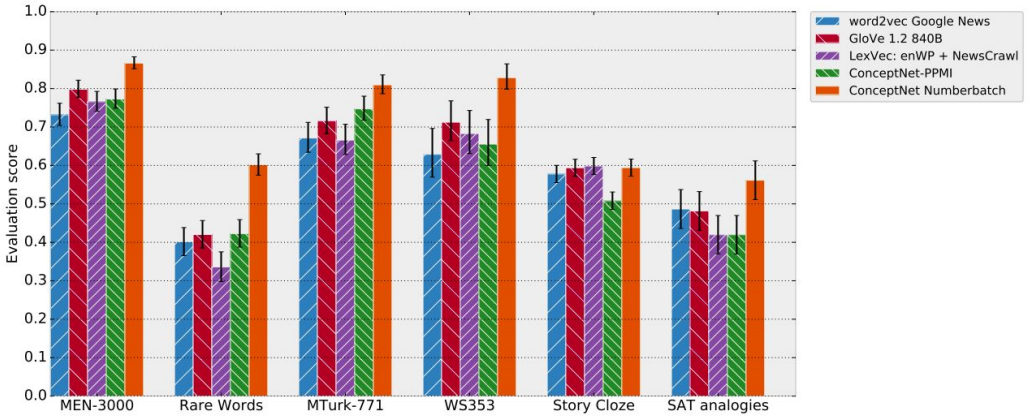


Figure 2: Performance of word embeddings across multiple evaluations. Error bars show 95% confidence intervals.

Como utilizar o ConceptNet?

- Documentação: https://github.com/commonsense/conceptnet5/wiki
- Web API: https://github.com/commonsense/conceptnet5/wiki/API

```
>>> import requests
>>> obj = requests.get('http://api.conceptnet.io/c/en/example').json()
>>> obj.keys()
dict_keys(['view', '@context', '@id', 'edges'])

>>> len(obj['edges'])
20

>>> obj['edges'][2]
{'@id': '/a/[/r/IsA/,/c/en/example/n/,/c/en/information/n/]',
 'dataset': '/d/wordnet/3.1',
 'end': {'@id': '/c/en/information/n',
  'label': 'information',
  'language': 'en',
  'sense_label': 'n',
  'term': '/c/en/information'},
 'license': 'cc:by/4.0',
 'rel': {'@id': '/r/IsA', 'label': 'IsA'},
 'sources': [{'@id': '/s/resource/wordnet/rdf/3.1',
   'contributor': '/s/resource/wordnet/rdf/3.1'}],
 'start': {'@id': '/c/en/example/n',
  'label': 'example',
  'language': 'en',
  'sense_label': 'n',
  'term': '/c/en/example'},
 'surfaceText': [[example]] is a type of [[information]]',
 'weight': 2.0}
```

Details about the data

- Languages - The hundreds of languages (and language families) represented in ConceptNet.
- Edges - The basic units of knowledge in ConceptNet and what they represent.
- Relations - Each edge has a relation. These are the kinds of relations we're using so far, and what they mean.
- URI hierarchy - What a URI such as `/s/contributor/omcs/dev` means.

Running your own copy

- Running your own copy - How to run your own copy of ConceptNet by reproducing its infrastructure and data, using Amazon Web Services
- Build process - How to build the ConceptNet data locally on a Unix system, and what you get as a result.
- Copying and sharing ConceptNet - ConceptNet is open data, and we want you to copy it and put it to new uses.
- Downloads

Como utilizar o ConceptNet Numberbatch?

- Documentação:
  - https://github.com/commonsense/conceptnet-numberbatch

Downloads

ConceptNet Numberbatch 19.08 is the current recommended download.

This table lists the downloads and formats available for multiple recent versions:

Version	Multilingual	English-only	HDF5
19.08	numberbatch-19.08.txt.gz	numberbatch-en-19.08.txt.gz	
17.06	numberbatch-17.06.txt.gz	numberbatch-en-17.06.txt.gz	17.06/mini.h5
17.04	numberbatch-17.04.txt.gz	numberbatch-en-17.04b.txt.gz	17.05/mini.h5
17.02	numberbatch-17.02.txt.gz	numberbatch-en-17.02.txt.gz	
16.09			16.09/numberbatch.h5

The .txt.gz files of term vectors are in the text format used by word2vec, GloVe, and fastText.

The first line of the file contains the dimensions of the matrix:

```
9161912 300
```

Each line contains a term label followed by 300 floating-point numbers, separated by spaces:

```
/c/en/absolute_value -0.0847 -0.1316 -0.0800 -0.0708 -0.2514 -0.1687 -...
/c/en/absolute_zero 0.0056 -0.0051 0.0332 -0.1525 -0.0955 -0.0902 0.07...
/c/en/absoluteless 0.2740 0.0718 0.1548 0.1118 -0.1669 -0.0216 -0.0508...
/c/en/absolutely 0.0065 -0.1813 0.0335 0.0991 -0.1123 0.0060 -0.0009 0...
/c/en/absolutely_convergent 0.3752 0.1087 -0.1299 -0.0796 -0.2753 -0.1...
```

The HDF5 files are the format that ConceptNet uses internally. They are data tables that can be loaded into Python using a library such as `pandas` or `pytables`.

The "mini.h5" files trade off a little bit of accuracy for a lot of memory savings, taking up less than 150 MB in RAM, and are used to power the ConceptNet API.

Exemplos de uso - chatbot e stance classification

A Deep Learning Model of Common Sense Knowledge for Augmenting Natural Language Processing Tasks in Portuguese Language - Carvalho et al (2020) - PROPOR

Comment

Tweet: Police reports released this morning indicate Mike Brown was a suspect in a "strong-armed" robbery in #Ferguson

Reply: It's a little scary to know they shoot and kill people suspected of crimes in Ferguson. @DIANAZOGA @AntonioFrench

Query

Tweet: UPDATE: 13 people being held hostage in #Sydney shop, Opera House evacuated after suspicious package found: reports http://t.co/n4D3yGjso9

Reply: @FoxNews why are you not airing the situation in Sydney ?

Deny

Tweet: MORE: Police believe three gunmen were involved in shootings in Ottawa this morning and are looking for two shooters. http://t.co/IHV4galEtC

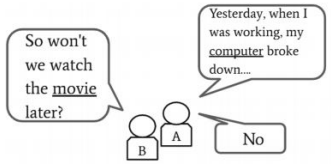
Reply: @CBSNews That's not what they just said at the Press Conference!

Support

Tweet: UPDATE: #CharlieHebdo's editor-in-chief was killed in attack - Magazine's lawyer http://t.co/kLKwMSlo7g

Reply:@SputnikInt May he rest in peace.

Fig. 5. Examples of the stance classification dataset.



- Computer is used to work ("Computador é usado para trabalhar", in Portuguese)
- Computer is used to watch movies ("Computador é usado para assistir filmes", in Portuguese)

Fig. 1. Dialogue examples and common-sense representation.

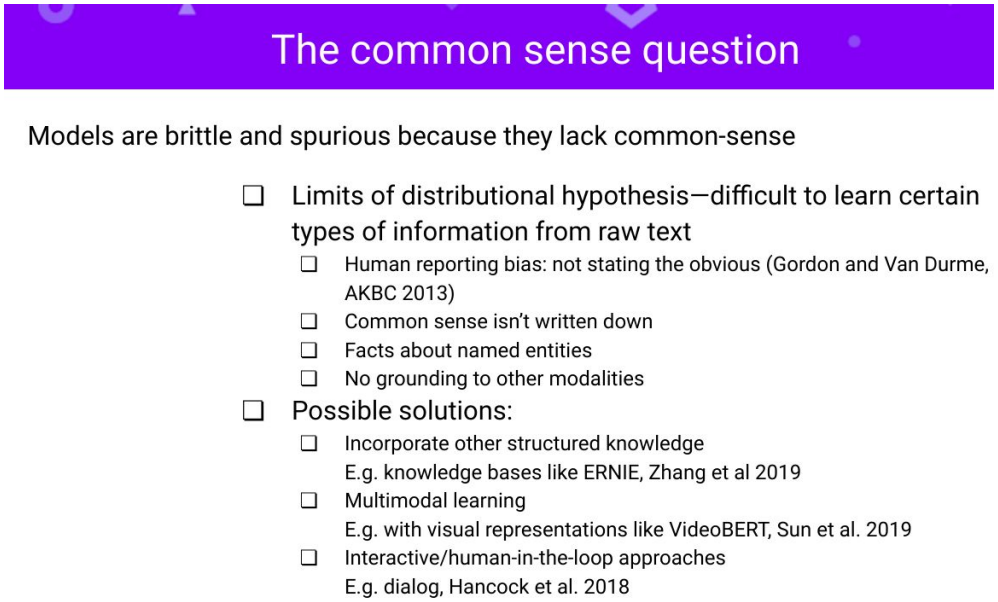
Table 3. Experiments Results with the use of DeepCS module in Chatbot application target.

Coherent		
	Scenario 1	Scenario 2
Recall	0.5389	0.6734
Precision	0.5682	0.5708
F1 measure	0.5644	0.5802

Table 2. Experiments Results with the use of DeepCS module in Stance Classification application target.

	Recall	Accuracy	F1 score	Comment	Deny	Query	Support
Scenario 1	0.6228	0.6228	0.3581	0.7660	0.1388	0.3406	0.1867
Scenario 2	0.6484	0.6484	0.3740	0.7883	0.1678	0.3870	0.1526

Exemplos de uso - commonsense pode melhorar a robustez de modelos



Slide 47: https://tinyurl.com/FutureOfNLP
Vídeo “Inductive Bias, Common Sense, Continual Learning in The Future of NLP” de HuggingFace: https://www.youtube.com/watch?v=cHiG0Ucv6Ck
Video “Commonsense Intelligence: Cracking the Longstanding Challenge in AI” https://livestream.com/newyorkacademyofsciences/nds2019-1

Fim