Neural Article Pair Modeling for Wikipedia Sub-article Matching

Muhao Chen[1], Changping Meng[2], Gang Huang[3], and Carlo Zaniolo[1]
[1]University of California, Los Angeles
[2]Purdue University, West Lafayette
[3]Google, Mountain View

Outline

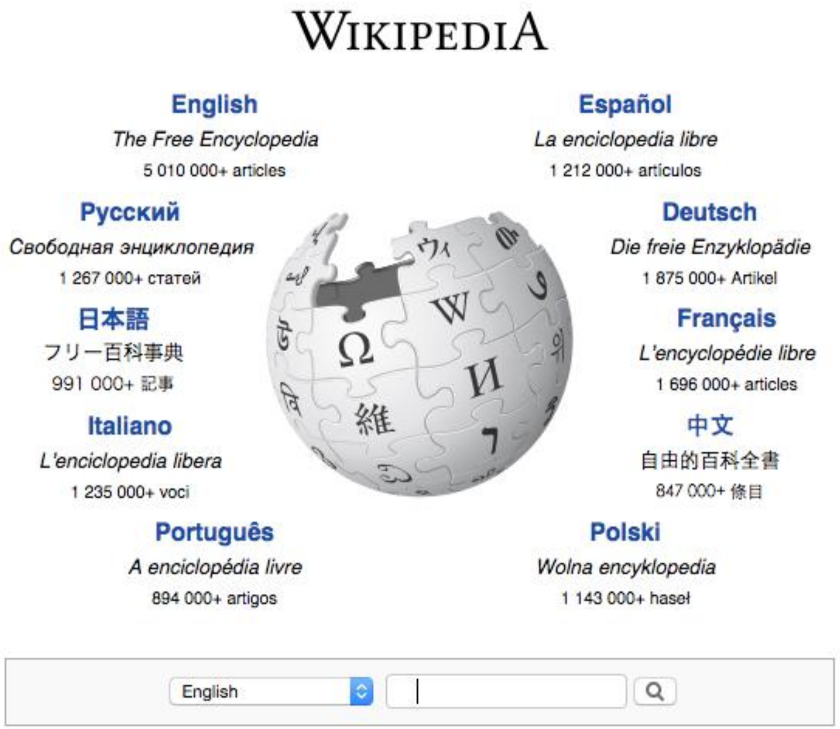
- Background
- Modeling
- Experimental Evaluation
- Future Work

Wikipedia: the source of knowledge for people and computing research

Essential sources of knowledge for people

- 45,567,563 encyclopedia articles
- 34,248,801 users

(As of 21 August 2018)



Countless knowledge driven technologies

- Knowledge bases
- Semantic Analysis
- Semantic search
- Open-domain question answering
- Named Entity Recognition
- etc.

Article-as-concept Assumption

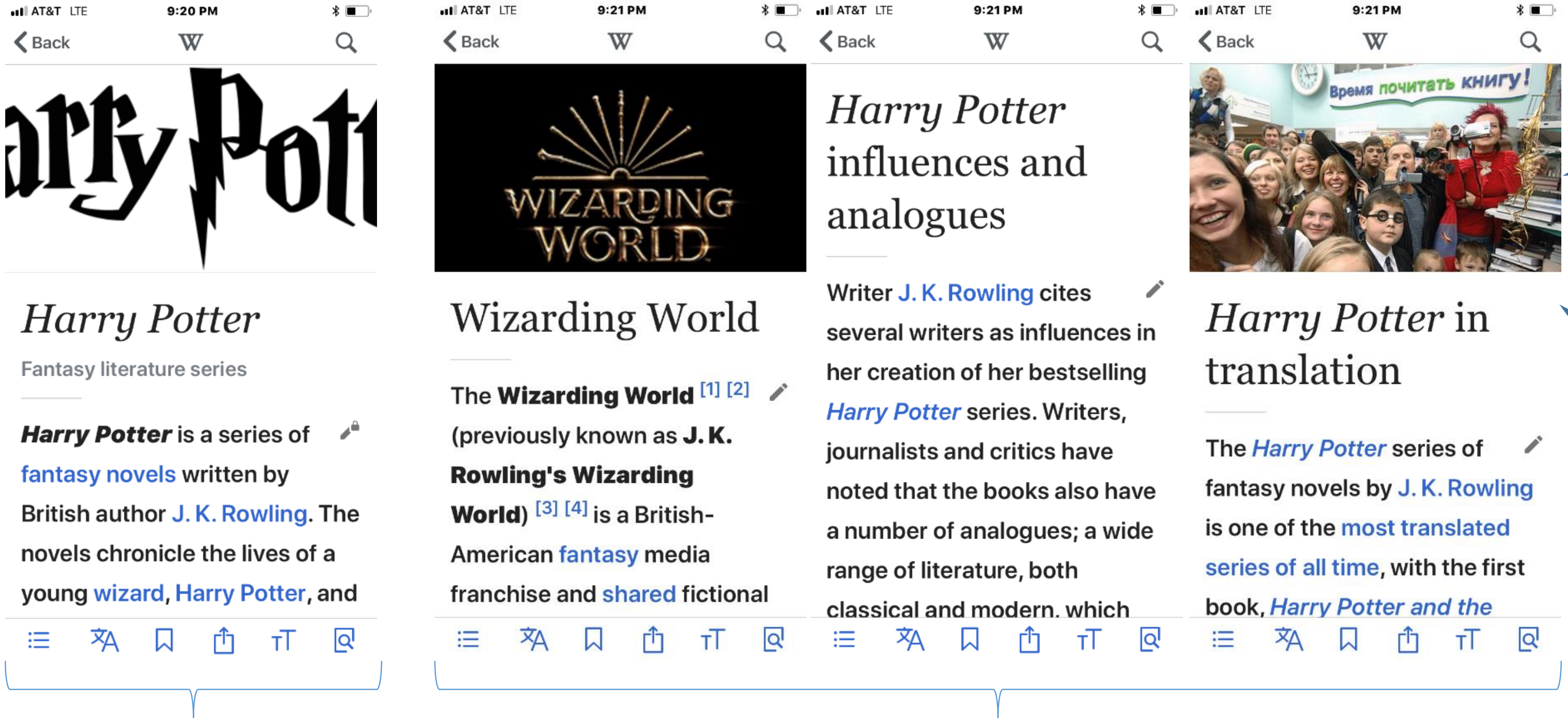
1-to-1 Mapping between entities and Wikipedia articles

Wikipedia-based computing technologies that rely on this assumption:

- Automated knowledge base construction
- Semantic search of entities
- Explicit and implicit semantic representations
- Cross-lingual Knowledge alignment
- etc.

Recent Editing Trends of Wikipedia

- Splitting different aspects of an entity into multiple articles.



Enhance human readability

Are problematic to Wikipedia-based technologies and applications

Main-article summarizes an entity. ***Sub-article*** comprehensively describes an aspect or a subtopic of the main-article.

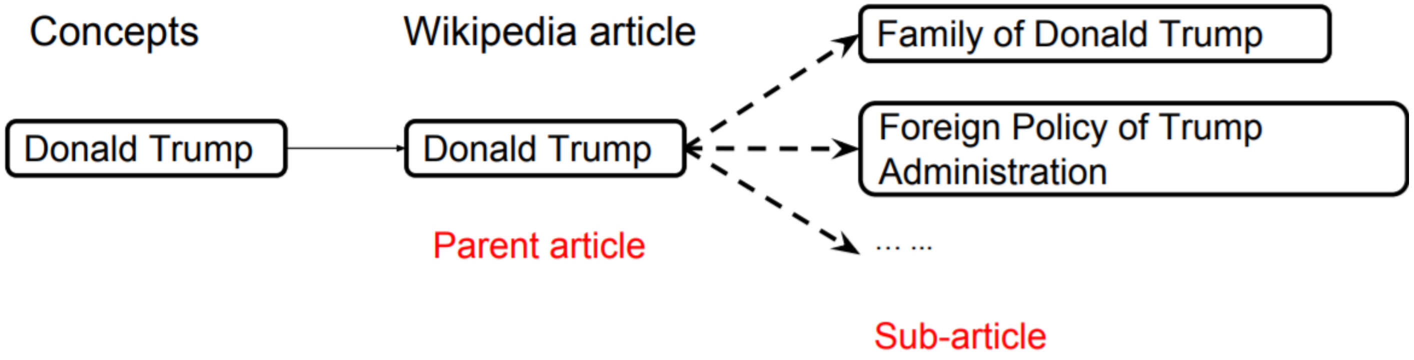
Violation of *Article-as-concept* Causes Problems to Existing Technologies

- Automated knowledge base construction: infoboxes and links are separated to multiple pages.
- Cross-lingual knowledge alignment and Wikification: one-to-one match does not hold.
- Semantic search: descriptions of entities are diffused
- Semantic representations: affected by the above
- …

We need to restore the scattered Wikipedia back

Problem Definition of Sub-article Matching

- Input: A pair of Wikipedia pages (A_i, A_j) (text contents, titles and links)



- Target: identify if A_i is the Sub-article of A_j
- Criteria of the sub-article relations:
  1. A_j describes an aspect or a subtopic of A_i
  2. The text content of A_j can be inserted as a section of A_i without breaking the topic of A_i

The sub-article relation conforms **anti-symmetry**.

Our Approach

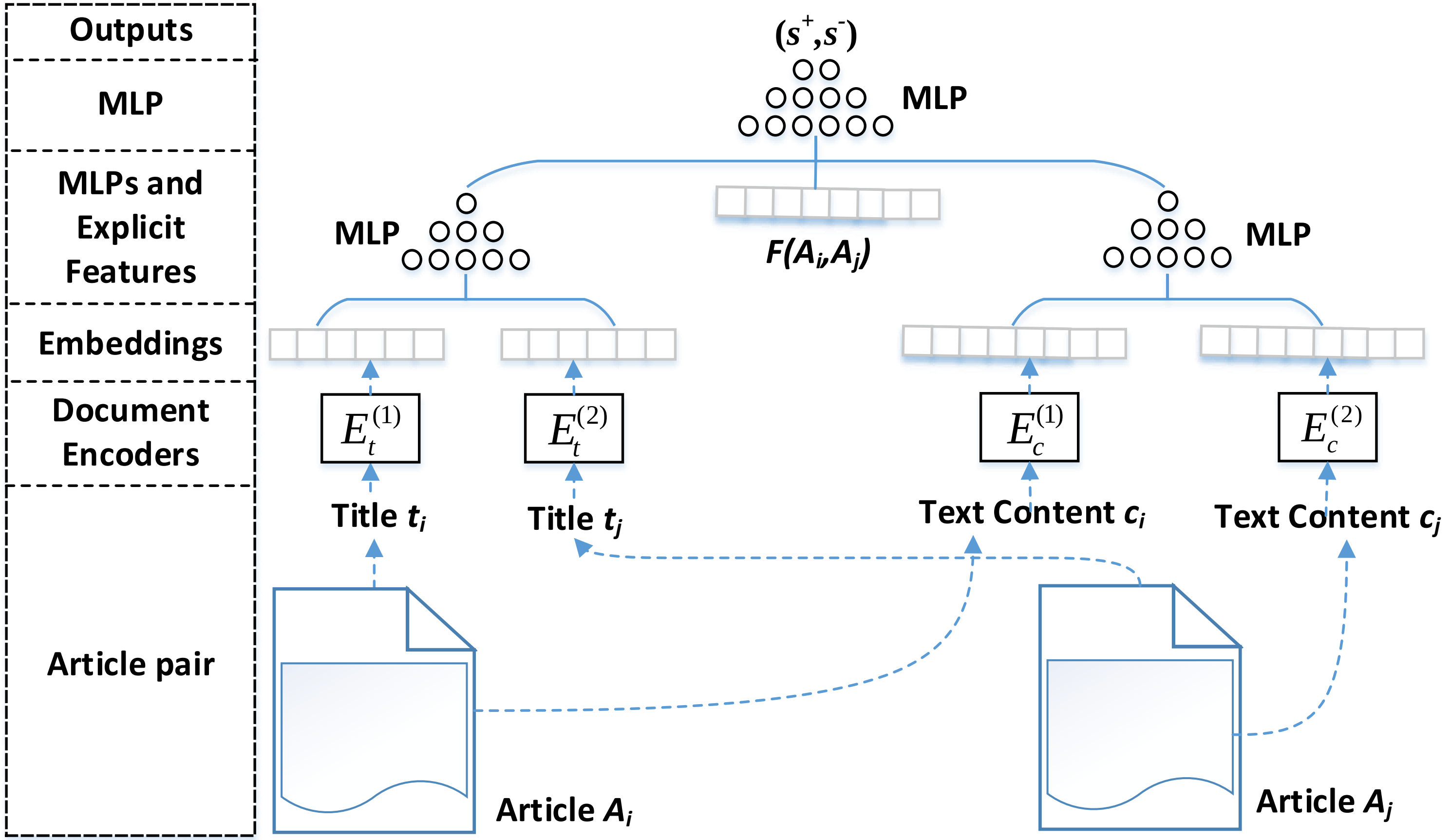
- A deep neural document pair model that incorporates
    1. Latent semantic features of articles and titles
    2. Comprehensive explicit features that measure the symbolic and structural aspects of article pairs

\- Obtains near-perfect performance on contributed data

\+ A scalable solution to extract high-quality M-S matching with thousand-machine MapReduce from the entire English Wikipedia.

\+ A large contributed dataset of 196k English Wikipedia article pairs for this task

Overall Learning Architecture

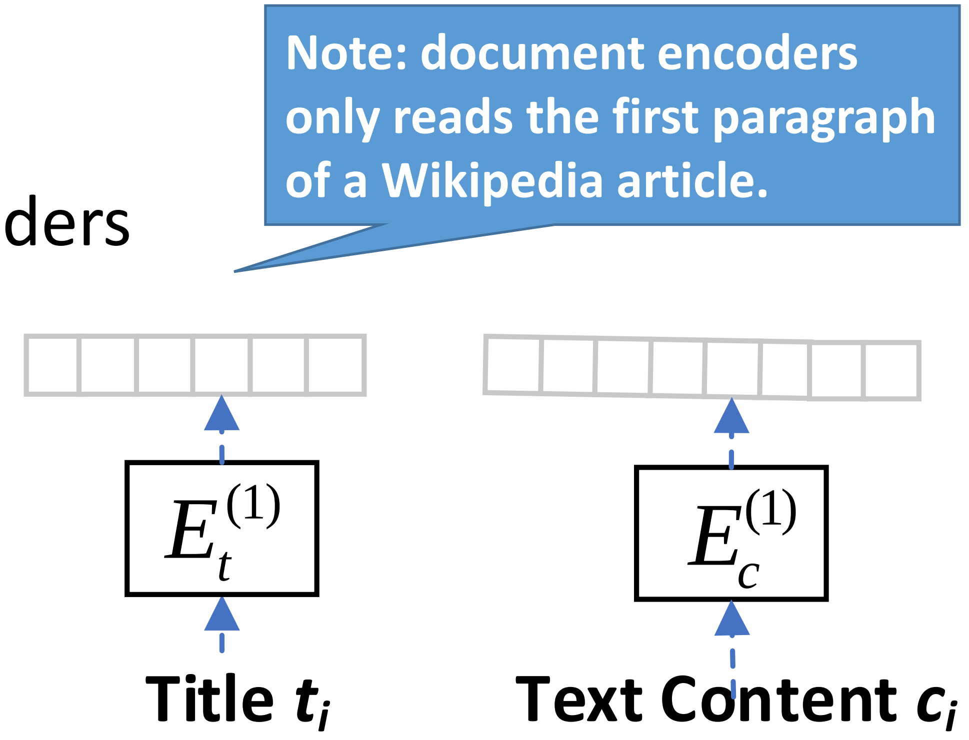


- Learning Objective: minimizes the binary cross-entropy loss

$$L = -\frac{1}{|P|} \sum_{p \in P} \left(l^+ \log s_p^+ + l^- \log s_p^- \right)$$

Neural Document Encoders

- Three types of neural document encoders
  1. CNN+Dynamic MaxPooling
  2. GRU
  3. GRU+Self-attention



- Word embedding layer: entity-annotated SkipGram

Explicit Features

r_{tto}	Token overlap ratio of titles.
r_{st}	Maximum token overlap ratios of section titles.
r_{indeg}	Relative in-degree centrality.
r_{mt}	Article template token overlap ratio.
f_{TF}	Normalized term frequency of A_i title in A_j text content.
d_{MW}	Milne-Witten Index.
r_{outdeg}	Relative out-degree centrality.
d_{te}	Average embedding distance of title tokens.
r_{dt}	Token overlap ratios of text contents.

Based on [Lin et al. 2017]

Additional

1. Symbolic similarity measures: r_{tto} r_{st} r_{mt} f_{TF} r_{dt}
2. Structural measures: r_{indeg} r_{outdeg} d_{MW}
3. Semantic measure: d_{te}

WAP196k—A Large Corpus of Main and Sub-article Pairs

1. Candidate sub-article selection → **2. Massive crowdsourcing** → **3. Negative cases generation**

Articles like ***German Army*** or ***Fictional Universe of Harry Potter***:

- Article titles that **concatenate two Wikipedia entity names** directly or with a proposition

- Annotators decide whether candidates from 1 are sub-articles. If so, find the corresponding main-articles.
- **Candidate article pairs** (positive and some negative matches) are selected based on **total agreement**.

Three rule patterns:

1. Invert positive matches.
2. Pair two sub-articles of the same main-article
3. Randomly corrupt the main-article of a positive match with an adjacent article.

Table 1: Statistics of the dataset.

#Article pairs	#Positive cases	#Negative cases	#Main-articles	#Distinct articles
195,960	17,349	178,611	5,012	32,487

1:10 positive to negative cases

Experimental Evaluation

- Task 1: 10-fold cross validation
  - Metrics: *Precision, Recall* and *F1* for identifying **positive cases**
- Baselines and model variants
  1. Statistical classification algorithms based on explicit features: Logistic Regression, NBC, LinearSVM, DecisionTree, Adaboost+DT, Random Forest, kNN. [Lin et al. 2017]
  2. Neural document pair models with latent semantics only (CNN, GRU, AGRU)
  3. Neural document pair models with latent semantics + Explicit feature (CNN+F, GRU+F, AGRU+F)

10-fold Cross Validation Results

Model	Explicit Features						
	Logistic	NBC	Adaboost	LinearSVM	DT	RF	kNN
Precision (%)	82.64	61.78	87.14	82.79	87.17	89.22	65.80
Recall (%)	88.41	87.75	85.40	89.56	84.53	84.49	78.66
F1-score	0.854	0.680	0.863	0.860	0.858	0.868	0.717

Model	Semantic Features			Model	Explicit+Semantic		
	CNN	GRU	AGRU		CNN+F	GRU+F	AGRU+F
Precision (%)	95.83	95.76	93.98	Precision (%)	**99.13**	98.60	97.58
Recall (%)	90.46	87.24	86.47	Recall (%)	**98.06**	88.47	86.80
F1-score	0.931	0.913	0.901	F1-score	**0.986**	0.926	0.919

- Semantic features are more effective than explicit features
- Incorporating both feature types reaches near-perfect performance

Feature Ablation Analysis

Table 3: Ablation on feature categories for CNN+F.

Features	Precision	Recall	F1-score
All features	99.13	98.06	0.986
No titles	98.03	85.96	0.916
No text contents	98.55	95.78	0.972
No explicit	95.83	90.46	0.931
Explicit only	82.64	88.41	0.854

Titles are then most important features (close to the practice of human cognition)

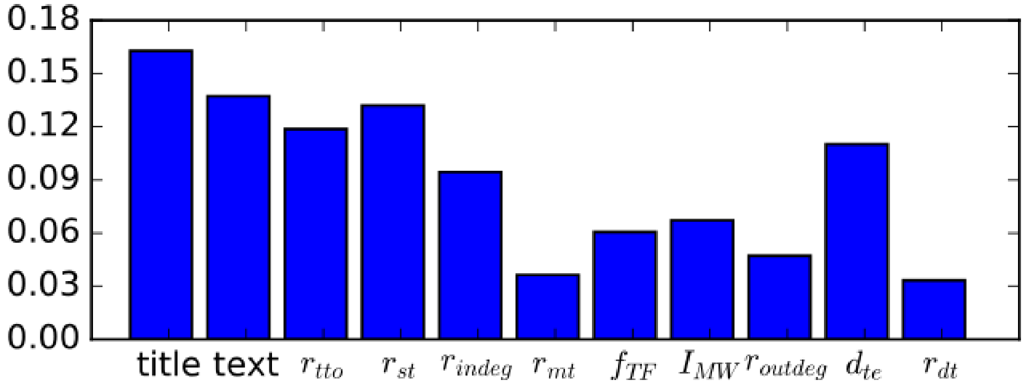


Fig. 3: Relative importance (RI) of features analyzed by Garson's algorithm. RI of each feature is aggregated from all folds of cross-validation.

Topological measures are relatively less important

Experimental Evaluation

- Task 2: large-scale sub-article relation mining from the entire English Wikipedia
- Model: CNN+*F* trained on the full WAP196k
- Candidate space: **108 million** ordered article pairs linked with at least one inline hyperlink
- Workload: ~ 9 hours with a 3,000-machine MapReduce

Extraction Results

- ~85.7% *Precision@200k*
- Avg 4.9 sub-articles per main-article

- Sub-article matching and Google Knowledge Graph

Table 4: Examples of recognized main and sub-article matches. The italicize sub-article titles are without overlapping tokens with the main article titles.

Main Article	Sub-articles
Outline of government	*Bicameralism*, *Capitalism*, *Dictatorship*, *Confederation*, *Oligarchy*, *Sovereign state*
Computer	Computer for operations with functions, Glossary of computer hardware terms, Computer user, *Timeline of numerical analysis after 1945*, Stored-program computer, Ternary computer
Hebrew alphabet	Romanization of Hebrew
Recycling by material	Drug recycling, *Copper*, *Aluminium*, Drug recycling
Chinese Americans	History of Chinese Americans in Dallas-Fort Worth, History of Chinese Americans in San Francisco, Anti-Chinese Violence in Washington
Genetics	Modification (Genetics), Theoretical and Applied Genetics, Encyclopedia of Genetics
San Marino	Economy of San Marino, San Marino national football team, Democratic Convention (San Marino), Banca di San Marino, Healthcare in San Marino, Flag of San Marino, Geography of San Marino
Service Rifle	United States Marine Corps Squad Advanced Marksman Rifle, United States Army Squad Designated Marksman Rifle
Transgender rights	LGBT rights in Panama, LGBT rights in the United Arab Emirates, Transgender rights in Argentina, History of transgender people in the United States, Transgender disenfranchisement in the United States
Spectrin	Spectrin Repeat
Geography	Political Geography, Urban geography, Visual geography, *Colorado Model Content Standards*
Nuclear Explosion	Outline of Nuclear Technology, International Day Against Nuclear Tests
Gay	*LGBT Rights by Country or Territory*, Philadelphia Gay News, Troll (gay slang), Gay literature
FIBA Hall of Fame	*Šarūnas Marčiulionis*
Arve Isdal	*March of the Norse*, *Between Two Worlds*
Independent politician	*Balasore (Odisha Vidhan Sabha Constituency)*
Mathematics	Hierarchy (mathematics), *Principle part*, Mathematics and Mechanics of Complex Systems, *Nemytskii operator*, *Spinors in three dimensions*, *Continuous functional calculus*, Quadrature, Table of mathematical symbols by introduction date, *Hasse invariant of an algebra*, Concrete Mathematics
Homosexuality	*LGBT rights in Luxembourg*, List of Christian denominational positions on homosexuality
Bishop	*Roman Catholic Diocese of Purnea*, *Roman Catholic Diocese of Luoyang*
Lie algebra	Radical of a Lie algebra, Restricted Lie algebra, *Adjoint representation*, Lie Group

Future Work

- Document classification
    1. Learning to differentiate main and sub-articles
    2. Learning to differentiate sub-articles that describe refined entities and those that describe abstract sub-concepts
- Extending the proposed model to populate the incomplete cross-lingual alignment

References

1. Lin, Y., Yu, B., Hall, A., & Hecht, B. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In *CSCW*. ACM 2017
2. Chen,M.,Tian,Y.,etal.:Multilingualknowledgegraphembeddingsforcross-lingual knowledge alignment. In: IJCAI (2017)
3. Chen, M., Tian, Y., et al.: Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In: IJCAI (2018)
4. Chen, M., Tian, Y., et al.: On2vec: Embedding-based relation prediction for ontology population. In: SDM (2018)
5. Dhingra, B., Liu, H., et al.: Gated-attention readers for text comprehension. In: ACL (2017)
6. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP (2014)
7. Jozefowicz, R., Zaremba, W., et al.: An empirical exploration of recurrent network architectures. In: ICML (2015)
8. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: CIKM (2008)
9. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: AAAI (2006)
10. Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." *IJCAI*. (2007)
11. Chen, Danqi, et al. "Reading Wikipedia to Answer Open-Domain Questions." ACL. (2017)

Thank You