Patent Citation Prediction with Seq2Seq

John Knowles, Sam Premutico
{jkn0wles, samprem} @ stanford.edu

Problem

Offer citation recommendations for patents in the US Patent Office

Explore embeddings over longer text documents

Designed as a multi-part system, utilizing NLP encodings, clustering techniques and supervised learning

Data / Task

Task

Represent document in vectorized form

Documents pruned to last 1000 words

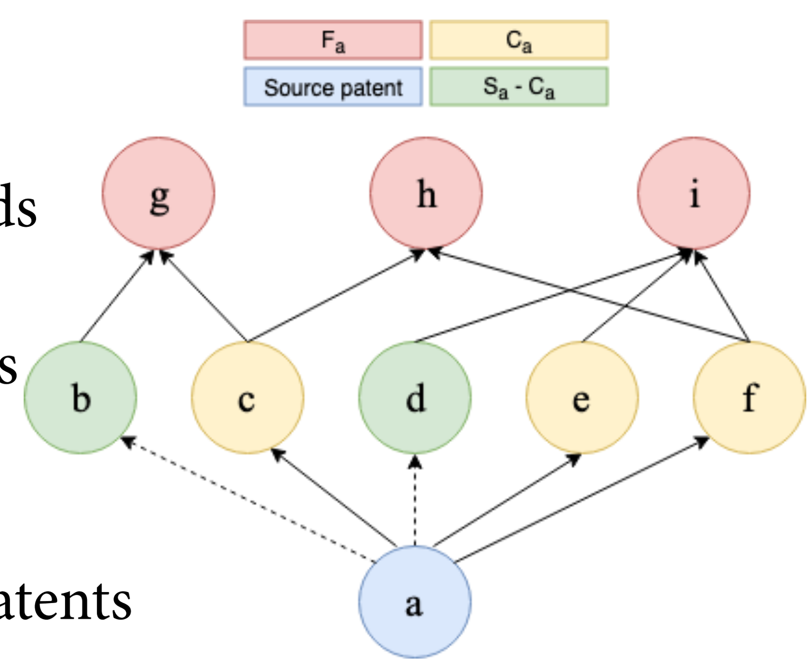
Vectorized into 256 components

Recommend citations using vector forms

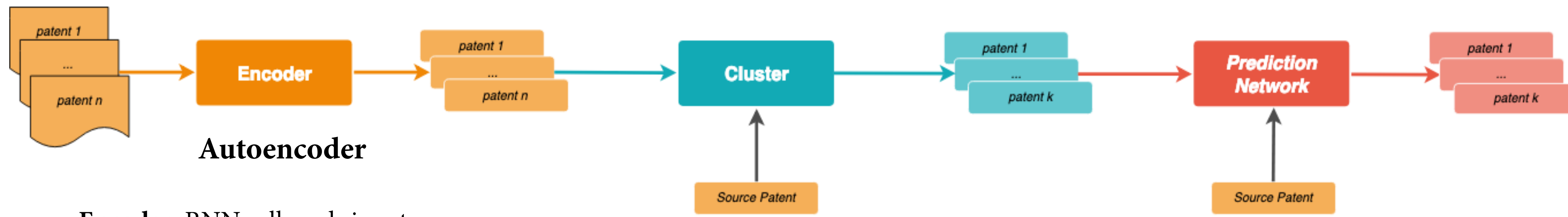
Framed as an clustering task

Data

train/test/dev : 75,000/5,000/5,000 patents



Approach



Autoencoder

Encoder: RNN cell reads input sequence, encodes to a 256 dimensional vectorized form

Decoder: RNN cell acting as a language model attempting to reconstruct the input sequence

Training: 10 hours on a NVIDIA Tesla M60 GPU using Microsoft Azure, using Cross Entropy loss with Logits and SGD optimizer

Clustering

Document cluster: top *K* documents based on cosine-sim to source patent

Constraint: output only citations from patents filed before the input patent

Cosine Similarity: $\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$

Prediction Network

Input: paired encoded source patent, encoded candidate patent

Output: Binary link prediction

Network: Apply convolutions to encodings to extract feature maps, feed through dense layers and output *score* through a sigmoid layer

Results

Table 1: **MRR Results Across K (cluster size)**

Model	K=50	K=100	k=300	k=500
TF-IDF Baseline	0.003	0.011	0.008	0.014
GloVe Baseline	0.014	0.016	0.016	0.018
Cluster Distance	**0.172**	**0.172**	0.174	0.175
Cluster + Predict	0.164	0.169	**0.185**	**0.185**

MRR (mean reciprocal rank): gives a score to an ordered list of possible responses to a query

Analysis

Autoencoder

- Performed poorly on the text reconstruction task (repetitions)
- Due to large inputs, we were not optimizing over text reconstruction, but rather quality of embeddings

Example Autoencoder Output:

"… application application **liquid thermal** …"

System

- Lower MRR patent recommendations contained large number of low degree seed patents. MRR was better overall across high-degree patents
- Prediction network did not boost results much beyond a simple clustering and results with no clustering performed poorly
- Able to significantly improve on baselines, which suggests that feature extraction from autoencoder provided meaningful representations

Conclusions/ Future Work

Conclusions

Large search spaces are **difficult**

Autoencoders provide a **general** way to extract meaningful feature representations

NLP modules can be used in contexts of **larger systems**

Future Work

Explore **deeper** neural architectures and different input representations for the autoencoder

Train over larger subset of the patent network

[1] I. Sutskever, O. Vinyals, Q. Le. Sequence to Sequence Learning with Neural Networks. 2014.
[2] G.Hinton, R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. 2006.
[3] R. Ying, Y. Li, X Li. GraphNet: Recommendation system based on language and network structure. 2017.