An Architecture for Scientific Document Retrieval Using Textual and Math Entailment Modules

Partha Pakray

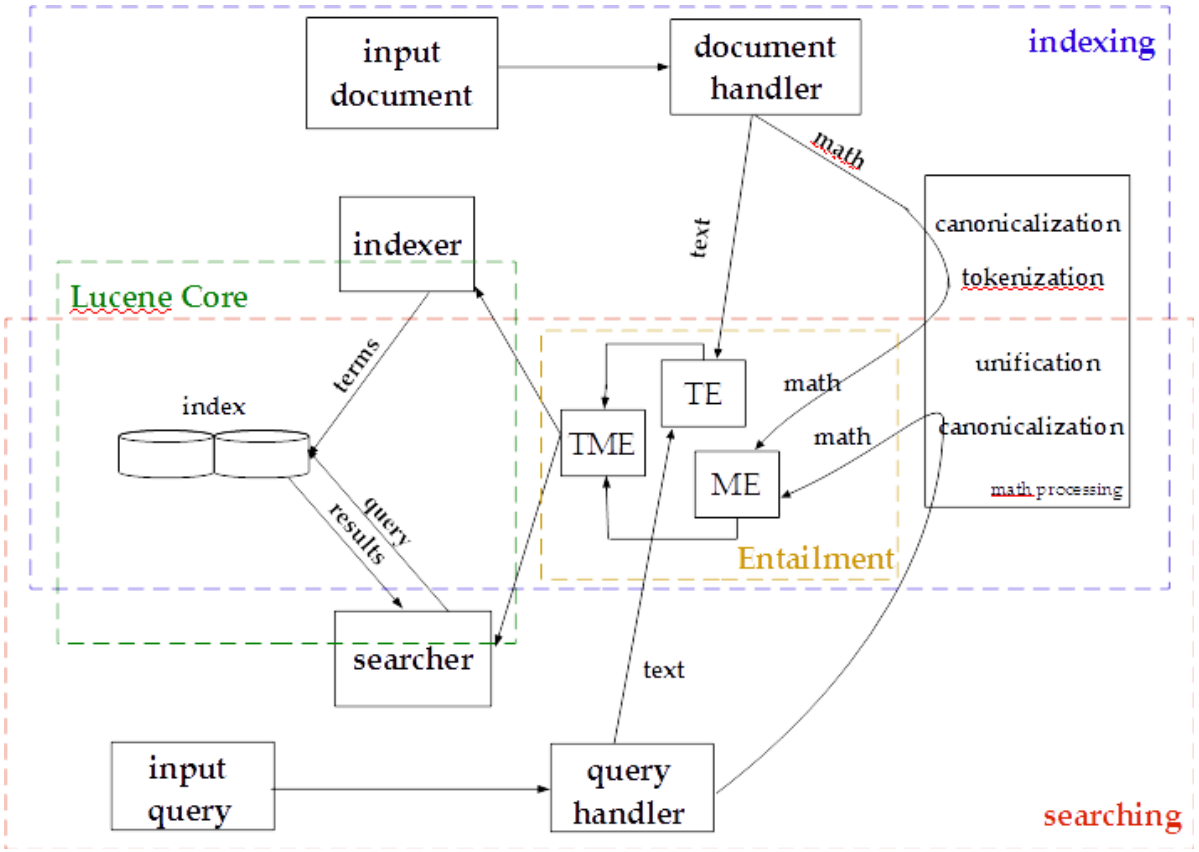
Petr Sojka

Objective & Goals

- Scientific document retrieval by using Entailment and Distributional Semantic Similarity

- Goal: to increase quality of retrieval (precision and recall) by handling natural language variations of expressing semantically the same in texts and/or formulae.

System Architecture



Semantic Similarity Measure

- Pre-trained word and phrase vectors of Google News dataset (about 100 billion words). The LSA word-vector mappings model contains 300-dimensional vectors for words and phrases.
- Gensim: Python framework for vector space modeling.
- Gensim for this experiment, and computed the cosine distance between vectors representing text chunks – sentences.
- Semantic Textual Similarity @ SemEval Tasks

Result

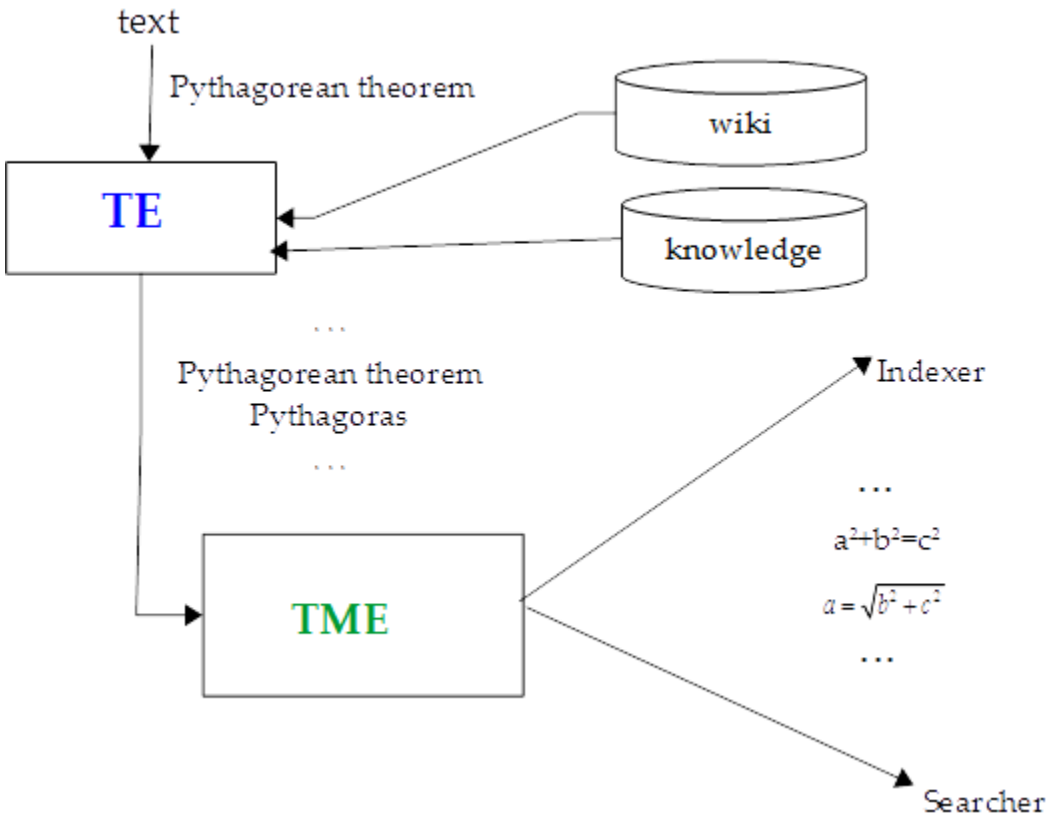
Table 1: SemEval-2014 Task 10: Multilingual Semantic Textual Similarity Test Result

Corpus	Winner score and team/run name		Our score
Deft-forum	0.5305	NTNU-run3	0.42812
Deft-news	0.7850	Meerakat_mafia-Hulk	0.67999
Headlines	0.7837	NTNU-run3	0.60985
Images	0.8343	NTNU-run3	0.71402
OnWN	0.8745	MeerkatMafia-paringWords	0.79135
Tweet-news	0.7610	DLS@CU-run1	0.76571

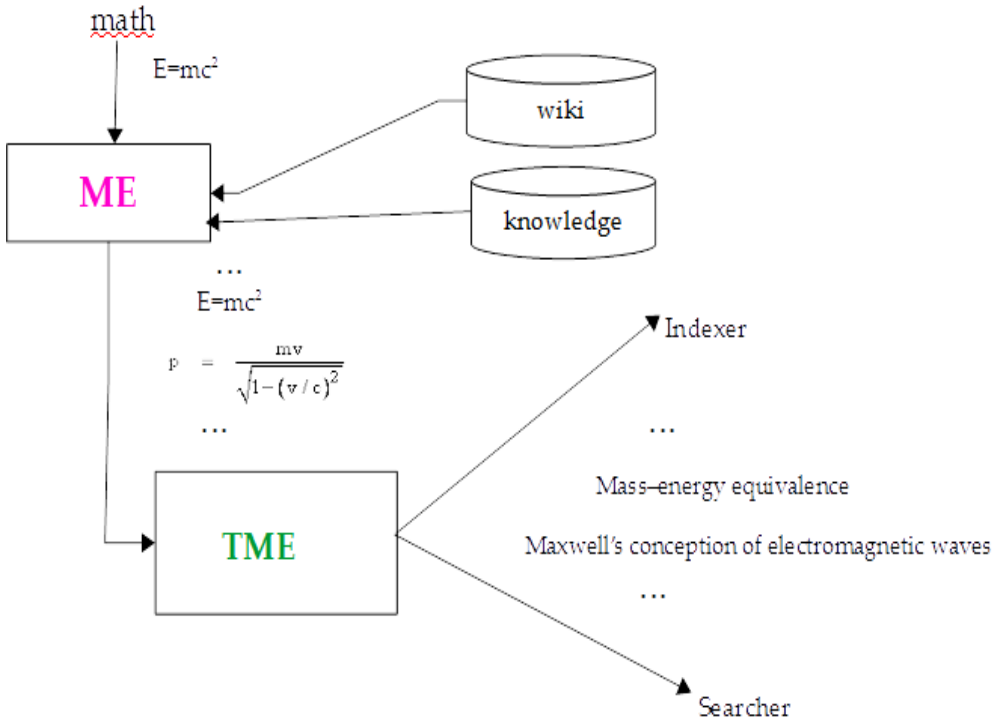
Table 2: SemEval-2013 Task 6: Semantic Textual Similarity Test Result

Corpus	Winner score and team/run name		Our score
Headlines	0.7838	UMBC_EBIQUITY-saiyan	0.62501
OnWN	0.8431	deft-baseline	0.71165
FNWN	0.5818	UMBC_EBIQUITY-ParingWords	0.38353
SMT	0.6181	UMBC_EBIQUITY-ParingWords	0.32951

TE & TME Module for Text



TE & TME Module for Math



Future Tasks

- We will build word and phrase vectors from Wikipedia articles.
- We will test on SemEval STS test data by using this generated vector from Wikipedia articles.
- We will also participate in STS evaluation track at SemEval 2015 Task.

thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
mersi
barka
welalin
tack
spas
kia ora
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
gracias
tapadh leat
хвала
asante
manana
obrigada
tenki
kiitos
dankie
dhanyavad
nanni
nandri
mauruuru
köszönöm
hvala
hvala
bedankt
bayarlalaa
gracje
enkosi
dziękuję
akun dankon
djiere dieuf
tau
mochchakkeram
mamnun
murakoze
chokrane
дякую
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
kam sah hamnida
rahmat
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
감사합니다
xièxie
ευχαριστώ
merci