CIKM, October 2016

Generalizing Translation Models in the Probabilistic Relevance Framework

Navid Rekabsaz[1], Mihai Lupu[1], Allan Hanbury[1], Guido Zuccon[2]

[1] TU WIEN, Vienna, Austria
[2] Queensland University of Technology, Brisbane, Australia

@NRekabsaz rekabsaz@ifs.tuwien.ac.at

@guidozuc g.zuccon@qut.edu.au

TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

Queensland University of Technology

Semantic Matching in IR

1. Matching by Query Reformulation
2. Matching with Translation Model
3. Matching with Term Dependency Model
4. Matching with Topic Model
5. Matching with Latent Space Model

Compared & Improved

Our contribution

Language Model: $score(q,d) = P(q|M_d) = \prod_{t_q \in q} P(t_q|M_d)$

[Ponte and Croft 1998]

Translation Language Model: $P(q|M_d) = \prod_{t_q \in q} \left(\sum_{t_d \in d} P_T(t_q|t_d) P(t_d|M_d) \right)$

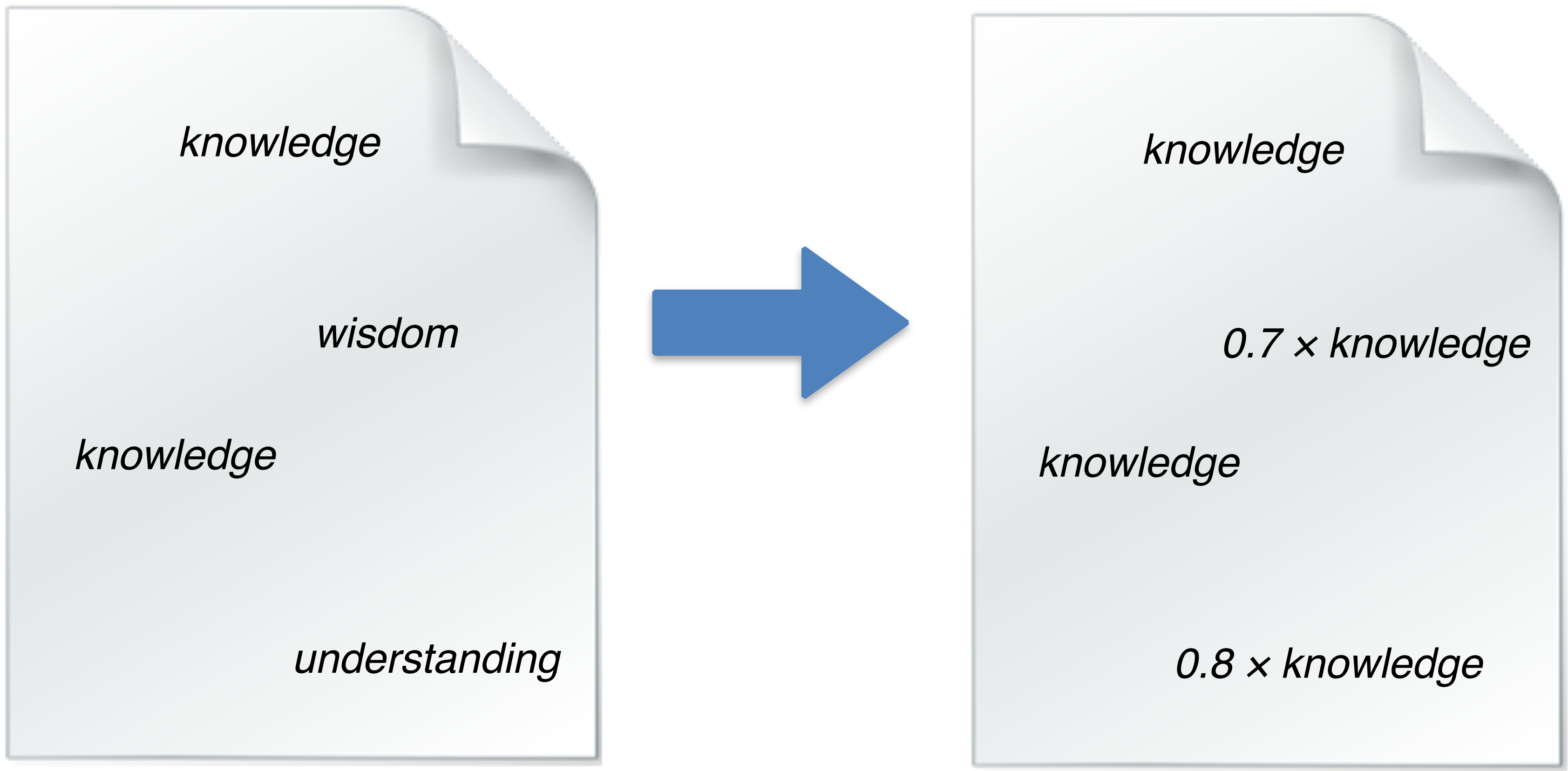
[Berger and Lafferi 1999]

What you will see…

- Generalization of the idea of translation models into Probabilistic Relevance framework models:
  - BM25
  - Pivoted Document normalization
  - BM25 Verboseness Aware
  - Multi-Aspect TF
- Integrating word embeddings into various IR models
- Comparing with Query Expansion (QE) methods, including Pseudo-Relevance Feedback (PRF)
- Experiments: significant improvements over original models and query expansion

Core Idea

Query: knowledge



- From term matching to (weighted) semantic matching
- We use the new documents for retrieval

Extended TF

Definition		Example
query term	t	*knowledge*
Word2Vec with threshold to select related words	$R(t)$	*understanding, wisdom, insight*
Cosine value for the translation probability	$P_T(t\|t')$	*{understanding: 0.8, wisdom: 0.7, insight: 0.65}*
normal term frequency	$tf_d(t)$	*= 2*
$\widehat{tf}_d(t) = tf_d(t) + \sum_{t' \in R(t)} P_T(t\|t') tf_d(t')$		*= 2 + 0.8 × 1 + 0.7 × 1 = 3.5*

Integrating into IR models

Pivoted Length Normalization (PL)

$$\Lambda(x) = log(1+x)$$

$$\sum_{t \in T_d \cap T_q} \frac{\Lambda(\Lambda(tf_d(t)))}{1 - s + s\frac{L_d}{avgdl}} tf_q(t) log \frac{|D|+1}{df_t}$$

Generalized Translation (PL GT)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{\Lambda(\Lambda(\widehat{tf}_d(t)))}{1 - s + s\frac{L_d}{avgdl}} tf_q(t) log \frac{|D|+1}{df_t}$$

$$\boxed{\widehat{df}_t = \left|\{d \in D : t \in T_d \vee \exists t' \in R(t), t' \in T_d\}\right|}$$

Extended Translation (PL ET)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{\Lambda(\Lambda(\widehat{tf}_d(t)))}{1 - s + s\frac{\widehat{L}_d}{\widehat{avgdl}}} tf_q(t) log \frac{|D|+1}{\widehat{df}_t}$$

Integrating into IR models

BM25

$$\sum_{t \in T_d \cap T_q} \frac{(k_1+1)\overline{tf}_d(t)}{k_1+\overline{tf}_d(t)} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{df_t+0.5}$$

BM25 Generalized Translation (GT)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{(k_1+1)\overline{\widehat{tf}}_d(t)}{k_1+\overline{\widehat{tf}}_d(t)} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{df_t+0.5}$$

BM25 Extended Translation (ET)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{(k_1+1)\overline{\widehat{tf}}_d(t)}{k_1+\overline{\widehat{tf}}_d(t)} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{\widehat{df}_t+0.5}$$

Integrating into IR models

The same procedure is applied to:

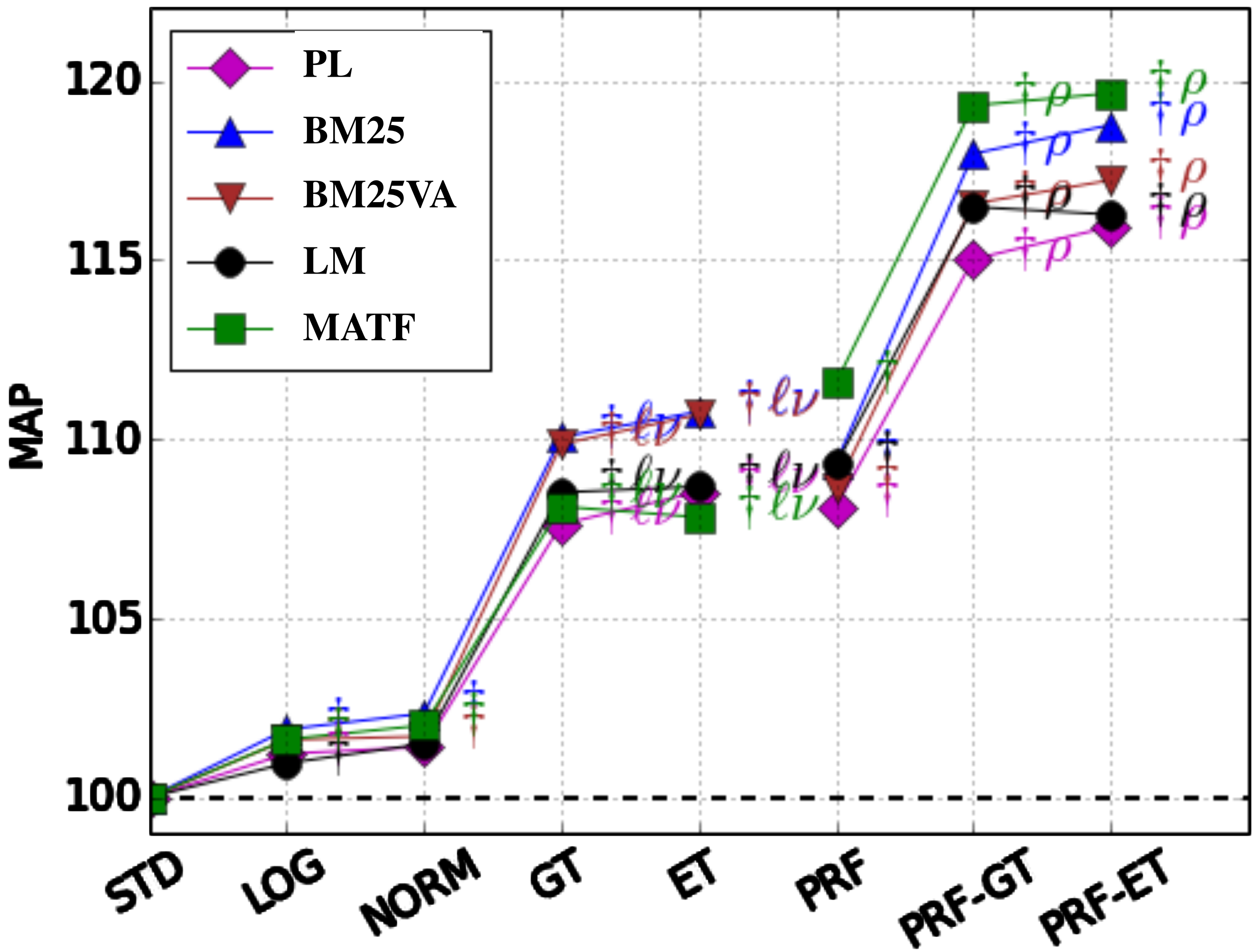
- BM25 Verboseness Aware [Lipani et al. 2015]
  - BM25VA GT
  - BM25VA ET
- Multi Aspect TF [Paik 2013]
  - MATF GT
  - MATF ET
- Language Model
  - LM GT ==> NTLM (TLM with word embeddings)[Zuccon et al. 2015]
  - LM ET

Experiments Setup

- Generalized and Extended Translation models: **GT** and **ET**
- Combining with Pseudo Relevance Feedback (**PRF**)
  - Translation models applied on results of PRF: **PRF-GT** and **PRF-ET**
- Baselines
  - **STD**: the original version of the models
  - **LOG**, **NORM**: two query expansion methods with word embedding
- Collections: TREC AdHoc 1-3, 6, 7, 8, HARD, CLEF eHealth 2015
- Evaluation metrics: MAP & NDCG@20

Results

Gain of MAP over original models, averaged over 6 collections



Stat sign. Test

† STD vs all

ν LOG vs GT,ET

ℓ NORM vs GT,ET

ρ PRF vs PRF-GT, PRF-ET

no sign b/w

GT vs PRF-GT

ET vs PRF-EF

Conclusion

- Generalization of translation model into various Probabilistic Relevance Framework models => significant improvements over original versions of the models
- Translation models have a complementary effect with PRF and combining them improves effectiveness

Questions?

Navid:
email me your feedbacks!

@NRekabsaz rekabsaz@ifs.tuwien.ac.at

@guidozuc g.zuccon@qut.edu.au

Extended …

Document frequency also adds the documents with related words

$$\widehat{df}_t = \left|\{d \in D\colon t \in T_d \vee \exists t' \in R(t), t' \in T_d\}\right|$$

The related terms are removed from the set of document terms

$$\widehat{T}_d = T_d \setminus \bigcup_{t \in q} \{t' \in R(t)\} \cup \{t \in q\colon R(t) \cap T_d \neq \emptyset\}$$ set of terms in document

$\widehat{L}_d = \sum_{t \in \widehat{T}_d} \widehat{tf}_d(t)$	document length
$\widehat{avgdl} = \frac{1}{\lvert D\rvert} \sum_{d \in D} \widehat{L}_d$	average document length
$\widehat{tf}_c(t) = \sum_{d \in D} \widehat{tf}_d(t)$	term collection frequency
$\widehat{L}_c = \sum_{t \in T} \widehat{tf}_c(t)$	collection size
$\widehat{avgtf}_d = \frac{1}{\lvert\widehat{T}_d\rvert} \sum_{t \in \widehat{T}_d} \widehat{tf}_d(t)$	average term frequency
$\widehat{mavgtf} = \frac{1}{\lvert D\rvert} \sum_{d \in D} \widehat{avgtf}_d$	mean average term frequency