Natural Language Retrieval of Grocery Products

Petteri Nurmi[1], Eemil Lagerspetz[1], Wray Buntine[2],
Patrik Floréen[1], Joonas Kukkonen[1], Peter Peltonen[3]

[1]Helsinki Institute for Information Technology HIIT, P.O. Box 68
FI-00014 University of Helsinki, Finland
firstname.lastname@cs.helsinki.fi
[2]National ICT Australia, 7 London Circuit, ACT 2601, Canberra, Australia
wray.buntine@nicta.com.au
[3]Helsinki Institute for Information Technology HIIT, P.O. Box 9800
FI-02015 Helsinki University of Technology TKK, Finland
peter.peltonen@hiit.fi

ABSTRACT

In this paper we describe modifications to a natural language grocery retrieval system, introduced in our earlier work. We also compare our system against an off-the-shelf retrieval tool, and show that our system is significantly better for top-ranked retrieval results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models, search process*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Design, Experimentation

Keywords

Information Retrieval, User Evaluation, Domain Specific IR

1. INTRODUCTION

In [2] we have introduced a natural language grocery retrieval system. Our initial evaluation resulted in a precision of 80% at rank one, and a mean average precision of c. 70%. In this paper we first describe modifications to our initial system (Sec. 2), after which we compare our system against Lemur, a representative off-the-shelf information retrieval tool. Our results indicate significantly improved performance for top ranks (one and two), and equal performance at later ranks.

2. MODIFICATIONS TO OUR SYSTEM

Compound Splitting and Index Expansion

The main change to our initial system is that we currently utilize compound splitting. Compound splitting is used in the indexing phase to generate new index terms, and in the query phase to determine whether the original query should be expanded with split constituents. The algorithm that we use for compound splitting has two phases. First we generate all valid split suggestions. For this we use an adaptation of the dictionary-based algorithm suggested in [1]. After the possible split suggestions have been generated, in the second phase we determine which split suggestion(s) to use. We make this decision by ranking the split suggestions and returning the original compound word and the suggestion with the highest score. As the ranking criterion we use the geometric mean of the popularity of the constituents.

Query Preprocessing

Products appearing in shopping lists often have case endings or are in plural. In Finnish, case and plural endings can cause the base stem of a word to change. The initial version of our retrieval system used stemming, but as many relatively common queries failed (especially for queries in plural form, e.g., berries or vegetables), we are currently using a lemmatizer on the user input. We use a freely available Finnish lemmatizer[1]. After lemmatization, we apply the compound splitting algorithm described above on the inputs.

Rank Augmentation

We maintain separate indexes for category and product name, which allows us to query product names and category names separately. We combine results of the two queries using a weighted extension to BM25 [3]. More specifically, we set

$$\begin{aligned} n'_j &= dm_j + n_j \\ f'_j &= dc_j + f_j \end{aligned} \quad (1)$$

where d is a weight term, n_j and m_j correspond to the number of products and categories where term j appears, and terms f_j and c_j correspond to category and product frequencies of term j. In the experiments, we use $d = 2.0$.

3. EVALUATION

Experiment Setting

In order to evaluate our retrieval system, we conducted a

[1]We use the Malaga language analysis tool (`http://home.arcor.de/bjoern-beutel/malaga/`) with the Suomi-Malaga Finnish grammar (`http://joyds1.joensuu.fi/suomi-malaga/suomi.html`). [Pages retrieved on: 2008-08-09]

Copyright is held by the author/owner(s).
CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

	Our system		Lemur		
Rank	**PR**	**EVAL**	**PR**	**EVAL**	**S**
1	83.39%	283	72.64%	318	* * *
2	78.21%	280	69.52%	315	* *
3	73.29%	277	71.52%	309	
4	73.36%	274	72.46%	305	
5	70.59%	272	72.28%	303	
MAP	69.77%	2703	66.92%	3029	*

Table 1: Results of the evaluation (PR = precision, EVAL = number of subjective assessments). The stars in the S column indicate the level of significance: * = 0.05, * * = 0.01, * * * = 0.001

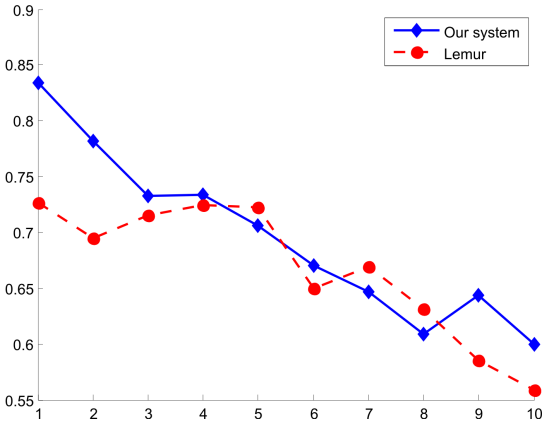

Figure 1: Plot of precision values at rank $1, \ldots, 10$. The solid line corresponds to our system, and the dashed line corresponds to Lemur.

user evaluation of the system. The evaluation was conducted as an intercept study at a large Finnish supermarket. The experiments spanned several days and different times of day. Hence, our participants should adequately represent the supermarket's customer base.

The evaluation was performed using a web interface that allowed to query for product names and give binary feedback to the results. As input, we used handwritten shopping lists, collected from the same supermarket. The results of different products were shown as consecutive ranked lists.

For each shopping list that was queried, we showed the top ten results for each item in the list. The participant was then asked to evaluate the accuracy of the results by clicking on a "thumbs up" or "thumbs down" image, located after each result. The participant was also allowed to leave an item unrated. We instructed the participants to rate the items based on how well they thought the items matched the query instead of considering whether they themselves would buy the corresponding product.

We also conducted the experiments using Lemur[2]. Lemur was used with the same preprocessing as our retrieval engine. Contrary to our system, Lemur does not have misspelling support and it uses only textual features for ranking; see [2] for details about our system. For both systems we used the same parameter values.

[2]`http://www.lemurproject.org/` [Retrieved: 2008-06-03]

Results

The evaluations resulted in a total of 2703 subjective relevance assessments for our system, and 3029 assessments for Lemur. From the results, we calculated the mean average precision (MAP) and precision measures at different ranks (i.e., P@N). The P@N values of the two systems are shown in Fig. 1, and the results are summarized in Table 1. The significance tests in the table were conducted using a t-test and assuming unequal variances. The degrees of freedom parameter for the test was estimated using the Welch-Satterthwaite approximation.

The results indicate that our system achieves around 84% accuracy at rank one. This is significantly better than the accuracy of Lemur, which is around 73%. At rank two, the difference between the systems is smaller, but remains significant. After rank three, the systems perform equally well. As the performance of Lemur is relatively stable between ranks one to five, this suggests that text-only retrieval can achieve around 70% accuracy in our domain. Our results also indicate that using category information and item popularities can lead to an increased accuracy in comparison to text-only retrieval.

The approximately linear decrease in the systems' performance is mainly caused by compound splitting and by the fact that certain queries are for a specific product. For many queries only few products are relevant, but compound splitting increases the number of retrieved products and hence also decreases accuracy at lower ranks. This problem occurs especially with berries, fruits, and vegetables, as they appear in various product names to indicate, e.g., flavor.

The accuracy of our system can be slightly improved by collecting more abbreviations, slang expressions and colloquialisms. However, beyond this, our results do not suggest any evident improvements. Nevertheless, we plan to experiment with a weighted ranking scheme that penalizes compound words.

Acknowledgments

This work was supported in part by the Finnish Funding Agency for Technology and Innovation TEKES, under the project Personalised Ubiservices in Public Spaces. This work was also supported in part by the IST Programme of the European Community, under the PASCAL network of excellence, IST-2002-506778. The publication only reflects the authors' views.

4. REFERENCES

[1] C. Monz and M. de Rijke. Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. In *Revised Papers from the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*, pages 1519–1541. Springer-Verlag, 2002.

[2] P. Nurmi, E. Lagerspetz, W. Buntine, P. Floréen, and J. Kukkonen. Product retrieval for grocery stores. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 781 – 782. ACM, 2008.

[3] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM CIKM*, pages 42 – 49. ACM, 2004.