Finding Substitutions of Rare Earth Elements Using Publication Data

Kirill Tumanov

Department of Knowledge Engineering, Maastricht University,
P.O. Box 616, 6200 MD Maastricht, the Netherlands

1 Introduction

Rare Earth Elements (REEs) attract a continuous interest of a wide range of studies and applications. Those elements possess specific chemical and physical properties beyond the characteristics of traditional elements [1, 2]. Nanotechnologies, being at the frontier of modern research are particularly eager to make use of REEs. However, since REEs are expensive to mine and process, developments in the field of their usage are limited.

Similar to the other industries, material scientists seek the productivity maximization of their compounds with the minimization of production costs. Therefore, a substantial effort is given to the search of more accessible substitutions of REEs. Obvious is a desire to preserve the unique properties of REEs.

It is clear that even with contemporary chemical and physical methods this search has a huge space of possible combinations. Thus, the studies on this topic rarely result in success. The aim of the present work is to demonstrate the automatic method facilitating the research in REEs' substitution finding.

In this study the possible substitutes are searched within publication data in a field of nanotechnologies. The idea is to identify the substitution candidates from the past research in order to give an insight of what elements should be tested in the current studies. This information should significantly benefit the researchers in the field by limiting the substitutions' search space.

2 Method Description

Substitutes can be found using abstracts as a comprehensive but condensed source of information about a publication. It can be assumed that the substitutes might have a specific relationship with any REE. Under this assumption abstracts are processed using modern Text Mining (TM) techniques [5].

The corpus of abstracts is preprocessed by a search of a verb phrase stem (VPS). When a stem is found, a sentence containing it is put away for the mining. A custom built TM system is then applied to the text fragments. The processing stages of the TM system include:

1. Sentence structure markup using StanfordNLP [3] (introduction of grammatical word-binding)
2. Co-reference resolution using ReVerb [4] (contextual enrichment of the sentence structures)
3. Relation extraction (such as a basic structure $\{\textit{Noun Phrase}\}$–$\{\textit{VPS}\}$–$\{\textit{Noun Phrase}\}$)

The output relations should finally be filtered based on their possessing of at least one REE. The assumption above should ensure that a candidate substitute is a part of such a relation. However, the scope of the corresponding substitutes is limited by the specified VPS.

The TM system may also be applied directly to the abstracts with the VPS of interest instead of the text fragments. This strategy results in a higher recall of substitutes, but in a lower precision. Note that as generally abstracts are kept short there is no critical increase of the processing time. The same holds for the whole article texts as well, but they require far more effort to process and are not always accessible. Even the increased potential of useful substitution discovery is not enough to use them instead.

3 Results

In this work a corpus included more than 190.000 abstracts of articles in nanotechnology published form 1999 to 2003. Around 550 text fragments with the stem "substitut" were collected. In the end around 70 candidate relations were extracted and after manual revision of an "expert" the final results were obtained. Some of them are shown in Table 1. Each "Substituent" is connected to a corresponding "Substance" by a generalized relation which may be read as "*used as a substitute in*". The elements and compounds of potential research interest are italicized.

Table 1: Exemplar substitutes of REEs and their chemical compounds

Substituent	Substance
Strontium	Lanthanum chromite Lanthanum manganites prepared by a sol-gel route Lanthanum *Cobalt* films prepared on *Platinum*
Titanium	Y_2Fe_{17} alloys
$Yttrium^{3+}$	*Zirconium*$^{4+}$
Molybdenum acid	Yttrium
Niobium	$Pr_8Fe_{86}B_6$ magnets
Neodymium	*Barium*
Samarium	$Nd_2Fe_{14}B$ *Barium*
Platinum	YPd_2B_2C

Similar techniques were applied to the abstracts directly as well. This resulted in an about 20% increase in the number of found substituents whilst the candidate substitutions number tripled and reached around 220 relations.

The results seem to be of high quality, but even so note that this information is only aimed to give a user an idea of a possible research direction, and it by no means may be used to claim that those related elements are direct substitutes and clearly not in all cases. To be able to grasp the context around the relations it is still necessary to read the whole corresponding article, where the found relation was extracted from.

4 Conclusion

It was demonstrated that the application of advanced TM techniques may successfully provide researchers with cues in the process of REEs substitutes identification. Although the quality of the substitutes still requires an expert evaluation, the dramatically decreased search space largely saves the expert work in this evaluation. The described framework may also be easily extended or modified to support extraction of other relations of interest in the field of nanotechnology as well as in alternative domains.

References

[1] Nomenclature of inorganic chemistry. *IUPAC Recommendations*, pages 51–52, 2005.

[2] Rare earths statistics and information. Technical report, United States Geological Survey, 2014.

[3] M. C. De Marneffe and C. D. Manning. Stanford typed dependencies manual. `http://nlp.stanford.edu/software/dependencies_manual.pdf`, 2008.

[4] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. *Proc. Conf. Empirical Methods in Nat. Lang. Proc.*, pages 1535–1545, 2011.

[5] R. Feldman and J. Sanger. *The text mining handbook*. Cambridge University Press, 2007.