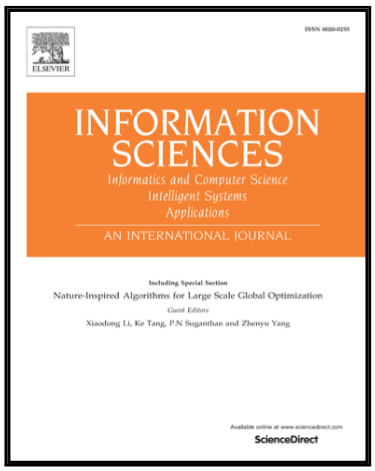


# Leveraging Linguistic Traits and Semi-Supervised Learning to Single Out Informational Content across How-to Community Question-Answering Archives

Daniel Palomera, Alejandro Figueroa

PII: S0020-0255(16)31680-2
DOI: 10.1016/j.ins.2016.11.006
Reference: INS 12613

To appear in: *Information Sciences*

Received date: 7 May 2016
Revised date: 1 October 2016
Accepted date: 16 November 2016

Please cite this article as: Daniel Palomera, Alejandro Figueroa, Leveraging Linguistic Traits and Semi-Supervised Learning to Single Out Informational Content across How-to Community Question-Answering Archives, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.11.006

# Leveraging Linguistic Traits and Semi-Supervised Learning to Single Out Informational Content across How-to Community Question-Answering Archives

Daniel Palomera, Alejandro Figueroa*

*Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andres Bello, Antonio Varas 860, Santiago, Chile*

**Abstract**

Community Question-Answering sites (e.g., Yahoo! Answers) have become large-scale knowledge bases of natural language questions formulated by their own members. In other to provide quick answers, these sites are compelled to make the best out of the content stored in their repository. Researchers have discovered that, on the one hand, many of these services are the confluence of an information-seeking and a social network that are constantly overlapping, and on the other hand, how-to questions are frequently published across these platforms. By and large, informational procedural questions are highly likely to expect informational answers, while non-informational manner questions target at socially interacting with other members of the community. In order to enhance user experience by reducing the delay in answering, these services are heartened to identify, retrieve and revitalize the content maintained in their knowledge bases. For this purpose, it is key to match the intent of new posted questions with the intention of archived answers that will be presented to the asker.

By manually annotating a reduced number of how-to questions and answers, we carried out an exploratory analysis that unveils a dichotomy between the interaction of these two networks. More precisely, we corroborate previous findings indicating that procedural questions are more likely to bear an informational goal, but our analysis is also extended to their answers, and it reveals that they exhibit a more conspicuous confluence. In substance, we find out that informational and non-informational answers are very likely to show up regardless of the end of the question. Then, we take advantage of this tagged set and of massive unlabelled material for exploiting two state-of-the-art single-view semi-supervised approaches aimed at discriminating informational from non-informational how-to content.

Moreover, our proposed models leverage assorted linguistically-motivated features, such as sentiment analysis and dependency parsing as well as named entity recognition.Our outcomes show that attributes, harvested from morphological and sentiment analysis, proven to be effective under a semi-supervised framework. At the expenses of low annotation costs, these linguistically-motivated semi-supervised models reached an accuracy of 84.25% and 74.41% for classifying questions and answers, respectively. In addition, we quantify the impact of automatically detecting informational/non-informational intents on the retrieval of best answers, i.e., an improvement of 4.12% in terms of precision at one.

*Keywords:* Semi-supervised learning; Question classification; Answer classification; Natural language processing; Knowledge base search; Community question answering;

## 1. Introduction

The most massively popular community Question Answering (cQA) services keep over 100 million resolved questions. As of December 2015, Yahoo! Answers[1] had about one hundred million members, and its knowledge base maintained more than four hundred million questions prompted since its inception (i.e, almost 250 million asked in English). As a logical consequence, cQA repositories are now perceived as vast sources of reusable information

*Corresponding author; phone: +56 (2) 27703795

*Email addresses:* d.palomera@uandresbello.edu (Daniel Palomera), alejandro.figueroa@unab.cl (Alejandro Figueroa)

[1] answers.yahoo.com