Wikipedia-based Context Visualisation

UC UNIVERSITY OF CANTERBURY Te Whare Wānanga o Waitaha CHRISTCHURCH NEW ZEALAND

Steve Manion
&
Dr Raazesh Sainudiin

Abstract

To be quoted out of *context* has the potential to be problematic for anyone, particularly in this day in age when our communications are often digitally recorded. Too often importance is placed on the *content* of what is stated, but not the *context* in which the statement is born from. Without recognising this, *content* can be used to state something completely different in a new *context*. In the field of Machine Translation this is a common cause of the underlying meaning and intention of a statement being lost in translation. To address this issue, the goal of this research is to anchor *content* to its original *context*, empowering translation algorithms to make more contextually sensitive decisions when selecting vocabulary for statements to translate, helping them to be understood as they were originally intended.

To do this we have extracted the context embedded within Wikipedia's *link* structure, which represents a large graph of topics (vertices) with *links* (edges) connected together when contributors collaboratively feel there is a significant relationship between two topics. The level of significance for each *link* varies according to its nature (such as direction, uniqueness), therefore we must apply a metric to reflect this so we can weight the edges of our graph and understand which *links* on a page are the most contextually significant. To achieve weighting we have used a metric involving Cosine Similarity along with a variation of TF-IDF [1], explained on the right side of this poster.

Example of Contextual Significance

General Context

We're sorry man, we have to cut off your leg.

Specific Context

We apologise sir, we must amputate your leg.

Fig. 1: The same sentence rewritten in 2 different contexts

In Fig. 1 above it is clear that in a hospital ward (i.e. speaking in a medical *context*) that the lower statement is contextually more appropriate. In everyday conversation ***cut off*** is more popular and versatile in its usage than ***amputate***, however if we can deduce that the *content* we are translating is from a medical *context*, our contextually weighted graph of Wikipedia can recommend ***amputate*** as a stronger candidate for the translation algorithm.

Presented on the 14th of September 2011 at the 18th Machine Learning Summer School in Bordeaux, France. Co-organized by INRIA and PASCAL.

Homonymous Case

Fig. 2: Topic cloud of 6 different contexts falling under one homonym

The cloud in Fig. 2 above illustrates six different competing *contexts*, all of which can be described using our single *query term* (which in this case is a homonym). Each *context* is represented by a different colour, furthermore each topic (vertex) has a font size representing its contextual significance with our *query term* (larger being stronger).

Q: What term (homonym) can we concatenate to cover all 6 contexts?

a) Wood _ _ _ _ b) Micro _ _ _ _ c) Potato _ _ _ _

d) Casino _ _ _ _ e) Chocolate _ _ _ _ f) Mod _ _ _ _

Synonymous Case

Fig. 3: Two different image collections representing two synonymous words

Both of these image collections in Fig. 3 above illustrate a different *context* however their *query terms* are synonymous with each other. Images that are larger and closer to the centre have stronger contextual significance with the image collection's *query term*.

Q: Both of these image collections represent 2 synonymous query terms given a particular context, can you guess each image collection's query term? (Hint: You can find them on this poster)

[1] Salton, G. and M. J. McGill (1983). Introduction to modern information retrieval. McGraw-Hill

[2] English Wikipedia XML dumps, http://dumps.wikimedia.org/enwiki/ (visited 28/05/2011)

[3]. Wikipedia mining for an association web thesaurus construction. In Proceedings of International Conference on Web Information Systems Engineering (WISE), pages 322-334, 2007 - **[Note the alternative approach to *link* direction treatment]**

Contextual Graph Construction

Initially we constructed our Wikipedia graph as a directed and unweighted entity from Wikipedia's XML dumps [2]. Then to meaningfully weight each *link* (edge) based on its contextual significance we took the following steps:

Step 1 – TF-IDF (Term Frequency – Inverse Document Frequency) is a metric to show how unique a *term* is in document relative to a whole collection of documents [1]. Its key benefit is that it reduces the significance of the common *terms* such as ***the*** whilst increasing the significance of unique *terms* such as ***amputate***. A TF-IDF vector of *terms* in a document characterises its *content* by highlighting the most significant *terms*. In our use of TF-IDF however we treat the inbound and outbound *links* of a document as our *terms* (hence LF-IDF), therefore identifying which *links* are the most significant based on *context* rather than *content*. Similar approaches have found this to be effective (for example see [3]).

Step 2 – Using LF-IDF vectors for each pair of linked documents, we can now weight each *link's* contextual significance based on Cosine Similarity. Cosine Similarity is a metric that returns a result between 0 and 1, with 0 meaning there is no similarity at all between two documents, and 1 meaning the two documents have exactly the same inbound and outbound *links*, thus are assumed to hold contextually equal *content*. Below are the results of using this method for the *query term* ***amputate***:

Table 1: Top 6 contextually significant topics for the query term Amputate according to Cosine Similarity scores calculated from LF-IDF vectors

Topic (Vertex):	Link (Edge) Weight
Traumatic amputation:	0.373171
Select Medical Corporation:	0.22833
Hemicorporectomy:	0.192546
Pulpotomy:	0.165739
Forequarter amputation:	0.115832
Team Zaryen:	0.100321

Fig. 4: Topic cloud for the query term Amputate