# Resultado de imagen de uocPEC 2: State of the art

Author: Teguayco Gutiérrez González

Studies: Master’s Degree in Data Science

## 2.1 Introduction

As explained in the introductory section, the problem of matching products coming from different sources brings with it the challenge of identifying identical products whose data may be missing (for instance, universal identifiers such as EAN or UPC codes could not be found) or presented in multiple formats. These difficulties make this matching task to be a non-trivial one and, as a result, a lot of research studies and tools have already arisen, most of them implementing Natural Language Processing (NLP) and Machine Learning techniques.

In this section, a comprehensive review of these previous works will be done based on the approaches they pose to analyse how they implement the solutions for this issue. It should also be noted that some of these works use more than one approaches.

## 2.2 Approaches

### 2.2.1 Text similarity

### 2.2.2 Attribute extraction

Most of the times, product titles include technical information which is crucial to identify them. However, in order to apply Machine Learning techniques, a more structured data about products should be managed. Therefore, the problem of obtaining these structured data about products from their textual information is arisen.

As an example, from the following product title:

# **Apple iPhone XS (64GB) – Gold**

It would be desired to filter it to obtain the following structured data:

{

“Brand”: “Apple”,

“ProductName”: “iPhone”,

“Model”: “XS”,

“MemoryCapacity”: “64”,

“MemoryCapacityUnit”: “GB”,

“Color”: “Gold”

}

This problem can be seen as an instance of an NLP task known as **Named-Entity Recognition** (NER) []. The goal in NER is to locate and classify the entities mentioned inside some given unstructured text.

#### 2.2.2.1 Simplest solution

One straightforward solution may consist in using **regular expressions** and **dictionaries** of known values for the attributes to extract. However, this approach can have several disadvantages.

On one hand, the use of regular expressions could lead to an inaccurate identification of attributes. For example, given the product name “Apple iPhone 32GB”, a regular expression can be written to extract the phrase “32GB”. However, this method would not be able to correctly resolve if this value correspond to RAM memory or to hard-disk capacity, as both attributes are commonly given in GB. Furthermore, as the information may be presented in many different formats, a regular expression should be written to handle every instance. For example, the inches of the screen for a certain TV could be written like “50-inches” or ‘50”’, so two different regular expressions are needed in this case.

On the other hand, the use of dictionaries may be a good approach for attributes with a fixed list of possible values like the colour. However, it could fail with attributes with an open list of possible values such as the brand, as new brands come eventually out on the market. In this latter case, the dictionary used to perform brand extraction should be regularly updated.

#### 2.2.2.2. Sequence labelling

Some studies like [1] and [2] address this issue by implementing more sophisticated methods. They first train some NER model to extract products features from their titles to then train Machine Learning models based on the structured data obtained to make the matching.

Thus, ***Conditional Random Fields*** (CRF) can be seen as the favourite alternative, as it is used by both of the above-cited works. It is a linear model for sequential labels. In fact, CRF can be considered as the sequential version of logistic regression.

Another implemented solution by [1] along with CRF is ***Structured Perceptron***, which is a supervised learning algorithm. By contrast, [2] goes beyond implementing CRF also with text embeddings.

The benefit of these sequence labelling algorithms is that they leverage the information given by the context of a given word in a certain product title. For example, in a product title like “Apple iPhone 4 (4GB RAM) - Black”, “4GB” could be recognized as RAM memory considering the next word is “RAM”.

Nevertheless, sequence labelling algorithms need a significant sizable training set (usually labelled using BIO encoding) whose manual labelling would require a huge usage time. Some implemented alternatives to avoid this manual task consist in the use of “distant supervision”, in which a training dataset is automatically built based on heuristics and rules.

### 2.2.3 Image recognition

Matching products identifying these by performing image recognition on them could constitute the hardest approach of all. It is not only the complexity of models like ***Convolutional Neural Networks*** (CNN), which are a common approach for image recognition problems, but also some other challenges that have to faced regarding products recognition from images: the same product can be found photographed from different perspectives, with different colours or different levels of brightness. Furthermore, CNNs could need a huge number of images to be trained, something that translates into managing and storing a lot of bytes of data.

## 2.3 References

<https://en.wikipedia.org/wiki/Named-entity_recognition>

AM2 (https://stats.stackexchange.com/users/13600/am2), Distant supervision: supervised, semi-supervised, or both?, URL (version: 2013-01-06): https://stats.stackexchange.com/q/46685