Injecting Multiple Psychological Features into Standard Text Summarisers

David E. Losada
Centro Singular de Investigación en Tecnoloxías
da Información (CiTIUS)
Universidade de Santiago de Compostela
david.losada@usc.es

Javier Parapar
Information Retrieval Lab
Department of Computer Science
University of A Coruña
javierparapar@udc.es

ABSTRACT

Automatic Text Summarisation is an essential technology to cope with the overwhelming amount of documents that are daily generated. Given an information source, such as a webpage or a news article, text summarisation consists of extracting content from it and present it in a condensed form for human consumption. Summaries are crucial to facilitate information access. The reader is provided with the key information in a concise and fluent way. This speeds up navigation through large repositories of data. With the rapid growth of online contents, creating manual summaries is not an option. Extractive summarisation methods are based on selecting the most important sentences from the input. To meet this aim, a ranking of candidate sentences is often built from a reduced set of sentence features. In this paper, we show that that many features derived from psychological studies are valuable for constructing extractive summaries. These features encode psychological aspects of communication and are a good guidance for selecting salient sentences. We use Quantitative Text Analysis tools for extracting these features and inject them into state-of-the-art extractive summarisers. Incorporating these novel components into existing extractive summarisers requires to combine and weight a high number of sentence features. In this respect, we show that Particle Swarm Optimisation is a viable approach to set the feature's weights. Following standard evaluation practice (DUC benchmarks), we also demonstrate that our novel summarisers are highly competitive.

CCS Concepts

•**Computing methodologies** → **Natural language processing;** •**Information systems** → *Content analysis and feature selection;*

Keywords

Summarisation, psychological features, LIWC, PSO

CERI '16 Granada, Spain

© 2016 ACM. ISBN 123-4567-24-567/08/06...$15.00

DOI: 10.475/123_4

1. INTRODUCTION

Automatic Text Summarisation (ATS) is indispensable for dealing with the rapid growth of online content. It is a powerful technology that can quickly digest and skim large quantities of textual documents. ATS has been employed in numerous application domains [13], such as news media –for instance, summaries of multiple stories on the same topic–, scientific literature –for instance, summaries of online medical literature–, or intelligence gathering -for example., biographical summaries for use by intelligence analysts. Automatic summarisers are also used prominently in web search to create summaries, also known as snippets, that are attached to search engine hits.

Extractive Summarisation has been an active subarea of ATS for decades. Extractive summarisers often apply shallow methods that extract salient parts of the source text and arrange them in some effective manner [13]. Features such as cue words, position within the text, or centrality (similarity to the text's centroid) are widely exploited for estimating salience. We constrain our discussion to sentence-based summarisers, which identify the most important sentences in the input and string them together to form a summary. In order to do so, three main tasks are performed: feature-based representation of every sentence, sentence scoring, and summary creation by sentence selection [14]. The first task involves a major language understanding challenge, but it is frequently addressed with simplified representational mechanisms (bag of words representations and frequency-based weighting schemes). More specifically, content-based features are computed by estimating how central the sentence's words are. And centrality is often measured from statistics such as term frequency (number of repetitions of the term in the input) and inverse document frequency, which uses corpus statistics to estimate how discriminative a term is.[1] Besides content-based features, other typical surface-level features are location features (sentences of greater topic centrality tend to occur in certain specifiable locations) or features based on cue words or phrases (for example, the phrase "in summary" could be a valuable indicator of a salient sentence).

In this paper we argue that language provides a full range of powerful indicators about emotions, cognition, social context, personality, and other psychological states; and these indicators can be exploited to extract salient sentences from text. Communication is not only about content. It is also about style and feelings. The style in which people use words reveals important aspects of their worlds, intentions, emo-

[1] This is the so-called tf-idf weighting scheme.

tional states and cognitive styles [19]. Verbal or written speech is a reflection of multiple factors. In the Social Sciences, the relationship between word use and many social and psychological processes has been actively studied. Psychometric properties of word use are informative about differences among individuals –for instance, men vs women–, about mental and physical health, and even about deception and honesty. Quantitatively analysing text also supplies a great deal of information about situational and social fluctuations.

Psychological word count approaches are potentially valuable for summarising text. When humans read text, the occurrence of certain psychological dimensions –for instance, positive or negative emotions, or cognition words– might be noteworthy. Besides content words that relate to psychological processes, linguistic style markers –for example, pronouns– are also known to yield unexpected insights. James W. Pennebaker, a renowned social psychologist and language expert, argues in his book entitled "*The Secret Life of Pronouns*" [18] that pronouns, prepositions and other common words are as distinctive as fingerprints; and analysing them is fruitful for a wide variety of applications. We intend to extend the scope of these psychometric studies to the area of Automatic Text Summarisation. It might be the case that the most salient or informative sentences in a document exhibit singular patterns of usage of psychological, social or linguistic elements. In this paper we try to shed light on this issue. Our study is an innovative way to understand what linguistic and psychological dimensions play a decisive role in revealing salient extracts of text. Our experiments are complementary to some analytical studies in the Social Sciences that have quantitatively evaluated different types of narrative and writing styles.

Combining standard and psycholinguistic features poses an optimisation problem. Dozens of sentence features have to be combined in order to tune the summarisers. Standard full search methods become unfeasible. We therefore propose and implement an optimisation strategy based on Particle Swarm Optimisation (PSO) [9]. PSO is a natural solution to this problem and our experiments show that the tuned weights work very well in practice.

The contributions of this paper are:

- We define novel sentence features for extractive summarisation based on Quantitative Text Analysis tools developed from Psychology theories. To the best of our knowledge, this is the first attempt to include this type of features in the area of Text Summarisation.

- We combine these psychologically-derived features with more standard sentence features (position, centroid and length). This leads to a weighting function for sentence scoring that aggregates multiple types of evidence.

- We employ PSO to estimate the optimal weights and show that this optimisation method is a robust alternative for tuning summarisers that have to combine a high number of sentence features.

- We inject this sentence scoring variant into a state-of-the-art summarisation system that produces non-redundant summaries of the desired size.

- We evaluate the resulting method for generic single-document and multi-document summarisation tasks, and compare it against standard baselines. Our new variant is very competitive and the relative performance of different feature sets gives insights into what psychological and linguistic dimensions are effective. We also analyse the types of summaries where our approach works the best.

- We analyse the feature's weights in the best performing summarisers and find interesting connections with studies in the Social Sciences about types of writing and analytical thinking. More specifically, our analysis reveals unknown linguistic characteristics of salient sentences and informs about stylistic and narrative elements that tend to occur in effective summaries.

2. SUMMARISATION METHOD

Natural language use has been linked to personality, social status, contextual behaviour, and other psychological factors [19]. The linguistic style of an individual reveals aspects of himself, his target audience, and the situation he is in. The way in which people use words –rather than the linguistic content– is a meaningful marker that has been studied with Quantitative Text Analysis methods. These methods statistically analyse the occurrence of standard grammatical units, psychologically derived categories, and other linguistic dimensions. This tracking of language is potentially valuable in Text Summarisation. For instance, the salient sentences of a text might exhibit certain stylistic patterns (for example, higher or lower percentage of personal pronouns; or higher or lower percentage of emotion words). Our method is based on exploiting these patterns in building extractive summaries.

Linguistic Inquiry and Word Count (LIWC) [23] is a Text Analysis Tool that computes the degree to which people use different categories of words. There are more than 70 LIWC categories, which are hierarchically organised. The complete list of categories and some examples are reported in [23]. The main top-level categories are linguistic processes (for example, personal pronouns, or verb tense), psychological processes (for example, affect words, emotions, or insights), and personal concerns (for example, work, or achievement). LIWC, which works from a dictionary of over 2300 words or word stems that have been associated to categories by independent judges, scans written text on a word by word basis and calculates the percentage of words in the text that match each category. LIWC is currently a reference tool that has been employed to quantitatively analyse a wide array of texts (for example, emails, speeches, or poems) in the context of numerous Text Analysis studies.

With LIWC, we computed sentence features to be taken into account for summarisation. From each category, we defined a sentence feature. For instance, the LIWC feature *we* represents the percentage of first personal plural pronouns –or alike– (for example, we, us, or our) in the sentence, and the LIWC feature *anger* represents the percentage of anger words (for example, hate, kill, or annoyed) in the sentence. We normalised all feature values to [0,1] range.

The interplay between these categories and different types of writing has been studied in the literature of the Social Sciences [18]. This led to findings such as the lower occurrence of I-words in formal writing, or the higher occurrence of quantifiers in analytical writing. LIWC categories are informative about writing styles and, therefore, potentially valuable to reveal salient content from text. This is precisely the main aim of our research. We intend to put LIWC categories into practice as a guidance mechanism for Text Summarisation. We are committed not only to designing innovative summarisation strategies but also to discovering what LIWC features are prominent in effective summaries. This latter objective will shed light on the (pyscho)linguistic constituents of abridged versions of text and will contribute towards understanding how information distillation works.

We estimated sentence salience by combining multiple types of evidence. Standard signals, such as the position of a sentence in a text, or the similarity between the sentence and the document's centroid, were combined with linguistically and psychologically derived signals obtained from LIWC. In areas like news summarisation the leading sentences of each document are known to provide much information about the document's contents. Therefore, extractive summarisers often weight the sentences appearing in the beginning of the documents more heavily. Another standard signal commonly employed in the literature of summarisation is centroid similarity. First, a centroid is computed for each document (or cluster of documents) to be summarised. This centroid uses standard statistics –for example, tf-idf weighting– to estimate which words are central to the document (or cluster of documents). Next, each source sentence is also represented as a vector of weighted words and matched against the centroid using the cosine similarity metric or some variant. This similarity weight promotes sentences whose overall resemblance to the whole document (or cluster of documents) is high.

Our aggregation method is based on linearly combining all feature weights (position, centroid and LIWC) and, next, the combined score is employed for ranking sentences. We incorporated this new sentence weighting method into a state-of-the-art summarisation system (MEAD).

MEAD [20] is a well-known toolkit that implements a variety of summarisation algorithms. Besides providing us with effective baseline summarisers, MEAD has a flexible and modular architecture that permits to incorporate new sentence features. First, a number of features are computed for each sentence of the document or cluster. Some built-in features are: position (the position of the sentence in the document), centroid (cosine overlap of the sentence with the centroid vector of the document or cluster), and length. All features range from 0 to 1.[2] Second, all feature values are linearly combined yielding an aggregated score for each sentence. These scores are used to build an initial ranking of sentences. Third, a re-ranking module removes sentences that are too similar to sentences already in the ranking. Finally, the resulting ranked set of sentences is used to produce a summary of the desired size.

3. EXPERIMENTS

We designed a complete pool of experiments to evaluate the usefulness of LIWC dimensions for selecting summary sentences. We worked with the following generic summarisation tasks from the Document Understanding Conferences (DUC)[3]: single-document summarisation (fully automatic summarisation of a single news article), and multi-document summarisation (fully automatic summarisation of multiple news articles on a single subject). Table 1 reports the main statistics of the datasets. Two datasets were used for parameter tuning (one for single-document summarisation and another one for multi-document summarisation) and the remaining datasets were used for test.

We implemented and tested the following summarisation algorithms:

- default MEAD. This is the default MEAD configuration based on centroid, position and length. The default feature weights are 1, 1, and 9, respectively. This means that sentences with less than 9 words are assigned an aggregated score of 0, while sentences with 9 or more words are assigned an aggregated score equal to the sum of its centroid and position scores.
- lead-based. This summariser takes the initial sentences of the document or cluster[4] to produce the summary.
- random. Random selection of sentences from the document or cluster.
- MEAD optimised (MEAD c+p tuned). This is the standard MEAD summariser with centroid, position and length but, rather than adopting 1 as default weight for centroid and position, we optimised the weights of these two features with the training collection.
- MEAD c+p+liwc. We computed LIWC features for each sentence and injected them into MEAD. The aggregated score was therefore a linear combination of centroid, position and the LIWC features. The combination weights were optimised with the training collection and the length cutoff was fixed to 9. Depending on the subset of LIWC features considered, this led to different summarisers (for example, MEAD c+p+liwc(ling.), MEAD c+p+liwc(psyc.), MEAD c+p+liwc (all)).

3.1 Parameter optimisation

A full exploration of the parameter space is not feasible for the MEAD c+p+liwc summarisers (80 dimensions in the case of MEAD c+p+liwc(all), 34 for MEAD c+p+liwc(psyc.), 23 for MEAD c+p+liwc(ling.), and 9 for MEAD c+p+liwc (pers.)). We applied Particle Swarm Optimisation (PSO) [9], which is a class of swarm intelligence techniques inspired by the social behaviour of bird flocking –or fish schooling– that runs a restricted search within the parameter space. PSO has been successfully employed for parameter tuning in many areas (for instance, in Information Retrieval [17, 4]). In this population-based stochastic method, the potential solutions, called *particles*, fly through the problem space following the current optimum particles. The movements of the particles are guided by the best known position of each particle in the search space as well as the entire swarm's best

[2] The first sentence of a document gets a position value of 1 and the remaining sentences are assigned linearly decreasing position scores. The length feature is treated as a cutoff. A sentence gets 1 if its length is above a given threshold and 0 otherwise.

[3] `http://duc.nist.gov`.

[4] In multi-document summarisation, the first sentence of each document will have the same scores, the second sentence of each document will have the same scores, and so forth; ties are resolved by the order of the documents in the definition of the cluster.

Table 1: Summarisation DUC datasets and tasks. The table reports the main statistics of the collections and how we used them in our experiments (train or test). All documents are news articles (from TREC) and the mean number of sentences per document is about 27.

Single-document summarisation					
		2001T (train split)	2001 (test split)	2002	
# docs		298	308	534	
summ. length		100 words	100 words	100 words	
train/test		train	test	test	
Multi-document summarisation					
	2001MT (train split)	2001M (test split)	2002M	2003M	2004M
# clusters	30	29	116	30	50
avg # docs/cluster	9.97	10.17	9.59	9.93	10
summ. length	100 words	100 words	100 words	100 words	665 bytes
train/test	train	test	test	test	test

known position. The process is repeated until a satisfactory solution is discovered.

The basic PSO algorithm is summarised in Algorithm 1. Each particle i stores its current position x_i^t, velocity v_i^t and its best known position pb_i^t at time t. The algorithm stores the best known position of the entire swarm (gb^t).

Algorithm 1 PSO basic Algorithm.

```
Initialise all particles $i$ with random positions $x_i^0$ in search space as well as random velocities $v_i^0$.
Initialise the particle's best known position ($pb_i^0$) to its initial position.
Calculate the initial swarm's best known position $gb^0$.
repeat
  for all Particle $i$ in the swarm do
    Pick random numbers: $r_p, r_g \in (0,1)$
    Update the particle's velocity: $v_i^{t+1} = a * v_i^t + b * r_p * (pb_i^t - x_i^t) + c * r_g * (gb^t - x_i^t)$
    Compute the particle's new position: $x_i^{t+1} = x_i^t + v_i^{t+1}$
    if $fitness(x_i^{t+1}) < fitness(pb_i^t)$ then
      Update the particle's best known position: $pb_i^{t+1} = x_i^{t+1}$
    end if
    if $fitness(pb_i^{t+1}) < fitness(gb^t)$ then
      Update the swarm's best known position: $gb^{t+1} = pb_i^{t+1}$
    end if
  end for
until termination criterion is met
return The best known position: $gb$.
```

In the algorithm, a, b and c are constants that separately control the importance of the three directions that determine the next velocity and position of the particle. The three components are usually referred as *inertia* (v_i^t), *personal influence* ($pb_i^t - x_i^t$) and *social influence* ($gb^t - x_i^t$). By updating the velocities with some element of randomness, the exploration of novel areas of the search space is enabled. This avoids stagnation in local minima and it is achieved by injecting random values, in the range (0,1), for the terms r_p and r_g. This yields a region of uncertainty around both positions, pb_i^t and gb^t.

PSO is an attractive solution for many optimisation or search problems. It has fewer parameters than popular simulated evolution methods, such as classic genetic algorithms [7]. Furthermore, with standard genetic algorithms, it is difficult to control the balance between exploration and exploitation. Even with low selective pressures there is a high probability that the population converges to a local optimum in few generations. Therefore, there is no guaranty that different potentially good areas in the space are simultaneously searched. Another alternative to deal with this problem comes from the niche and speciation techniques based on *fitness sharing* [7], or the distribution of the population in races or islands that evolve independently and simultaneously, which periodically interchange their best genetic material. PSO, in contrast, intrinsically explores the search space with a concentration of the population around the promising areas.

Using the standard PSO algorithm, we optimised the ROUGE-2 metric, which is explained next, with a population of 100 particles and used the recommended values for PSO [5].

3.2 Performance measures

We followed existing practice to automatically determine the quality of a summary by comparing it to summaries created by humans. ROUGE measures [12] are widely applied for evaluating automatic summaries. Among all ROUGE measures, ROUGE-2 and ROUGE-SU4 have shown to be correlated with human's judgements [22] and, therefore, we adopted them as our reference metrics. ROUGE measures are recall-oriented metrics that count the number of overlapping units (for example, n-grams) between the automatic summary and the manual summary. ROUGE-2 is based on counting bigrams and ROUGE-SU4 is a variant that allows for gaps (*skip-bigram* with maximum gap length of 4) and also considers unigrams.

3.3 Results

The performance of each summariser is reported in Table 2 (single-document summarisation) and Table 3 (multi-document summarisation). Not surprisingly, the random summariser is the weakest for both types of tasks. The lead-based summariser is comparable to default MEAD for single-document summarisation. In contrast, the lead-based summariser is inferior to default MEAD for multi-document summarisation. This makes sense. Summarising a single document is easier and we can benefit from the journalistic style of writing (main ideas first). But summarising multiple documents is harder and choosing leading sentences is problematic.

Optimising the centroid and position weights did not give much added value: default MEAD and MEAD c+p tuned

Table 2: Test results (Single-Document Summarisation). ROUGE-2 and ROUGE-SU4 scores are reported together with their 95% confidence intervals (in brackets). For each collection and performance measure the highest score is bolded.

	ROUGE-2	ROUGE-SU4
DUC2001		
default MEAD	.1793 (.1660,.1941)	.1813 (.1698,.1926)
random	.1277 (.1167,.1401)	.1420 (.1336,.1517)
lead-based	.1931 (.1796,.2071)	.1825 (.1726,.1934)
MEAD c+p	.1928 (.1792,.2067)	.1820 (.1721,.1927)
MEAD c+p+liwc(all)	.1918 (.1787,.2055)	.1848 (.1741,.1954)
MEAD c+p+liwc(ling.)	**.1953** (.1820,.2091)	**.1882** (.1777,.1992)
MEAD c+p+liwc(psyc.)	.1913 (.1775,.2054)	.18550 (.1744,.1969)
MEAD c+p+liwc(pers.)	.1919 (.1783,.2051)	.1865 (.1756,.1972)
DUC2002		
default MEAD	.1995 (.1912,.2080)	.1928 (.1855,.2000)
random	.1437 (.1357,.1520)	.1506 (.1441,.1573)
lead-based	.2067 (.1986,.2154)	.1928 (.1862,.2000)
MEAD c+p	.2069 (.1988,.2153)	.1931 (.1866,.1997)
MEAD c+p+liwc(all)	.2095 (.2015,.2182)	.1983 (.1919,.2053)
MEAD c+p+liwc(ling.)	**.2096** (.2014,.2180)	**.1984** (.1919,.2049)
MEAD c+p+liwc(psyc.)	.2110 (.2024,.2190)	.1964 (.1894,.2035)
MEAD c+p+liwc(pers.)	.2092 (.2011,.2180)	.1943 (.1872,.2016)

perform roughly the same. Including LIWC features was somehow beneficial, particularly for multi-document summarisation. In most of the cases, the MEAD c+p+liwc(ling.) summariser performed the best. Although the improvements over the baselines are modest (and, usually, statistically non-significant), the set of LIWC features associated to linguistic processes looks promising. The other subsets of LIWC features and the complete set of LIWC features (All) did not yield any consistent improvement. In the light of these results, we decided to further analyse the MEAD c+p+liwc(ling.) summariser, trying to understand when it is useful. This analysis is described in the next subsection.

3.4 Analysis of MEAD c+p+liwc(ling.)

The MEAD c+p+liwc(ling.) summariser is consistently better than the baseline summarisers. However, most of the improvements are not statistically significant. This suggests that rather than modestly improving all types of summaries, MEAD c+p+liwc(ling.) is working well for some cases but is degrading the performance for certain individual summarisation cases. For each summarisation case (document or cluster), we took the individual ROUGE-2 score of a baseline summariser (MEAD c+p tuned) as an estimator of the difficulty to summarise the document or cluster.[5] We computed the difference between the ROUGE-2 score of the MEAD c+p+liwc(ling.) summariser and the ROUGE-2 score of the baseline summariser. By plotting these differences against the baseline's ROUGE-2 scores, we can observe how MEAD c+p+liwc(ling.) behaves with varying degrees of difficulty. The plot (Fig. 1) is quite revealing. In all cases, MEAD c+p+liwc(ling.) has a tendency to work well for difficult summarisation cases (low ROUGE-2 scores) and to be harming for easier summarisation cases (high ROUGE-2 scores). This happens for both single-document and multi-document summarisation. Furthermore, the regression analysis of these two variables (X= ROUGE-2 baseline, Y= diff ROUGE-2 scores) concludes that there is a negative association between them. And we can reject the hypothesis that the slope is 0 (no association) in all cases (all p-values are less than .05).

Table 3: Test results (Multi-Document Summarisation). ROUGE-2 and ROUGE-SU4 scores are reported together with their 95% confidence intervals (in brackets). For each collection and performance measure the highest score is bolded.

	ROUGE-2	ROUGE-SU4
DUC2001M		
default MEAD	.0510 (.0374,.0646)	.0828 (.0682,.0986)
random	.0310 (.0213,.0424)	.0645 (.0544,.0747)
lead-based	.0303 (.0213,.0400)	.0639 (.0548,.0744)
MEAD c+p	.0579 (.0422,.0746)	.0888 (.0723,.1063)
MEAD c+p+liwc(all)	.0550 (.0422,.0693)	.0856 (.0730,.0997)
MEAD c+p+liwc(ling.)	**.0667** (.0499,.0824)	**.1004** (.0825,.1185)
MEAD c+p+liwc(psyc.)	.0586 (.0420,.0752)	.0867 (.0722,.1023)
MEAD c+p+liwc(pers.)	.0563 (.0396,.0750)	.0864 (.0708,.1040)
DUC2002M		
default MEAD	.0684 (.0610,.0769)	.0950 (.0870,.1032)
random	.0355 (.0301,.0413)	.0710 (.0659,.0764)
lead-based	.0433 (.0369,.0504)	.0659 (.0601,.0716)
MEAD c+p	.0610 (.0550,.0678)	.0963 (.0898,.1030)
MEAD c+p+liwc(all)	**.0720** (.0643,.0810)	.1006 (.09371,.1083)
MEAD c+p+liwc(ling.)	.0711 (.0637,.0789)	**.1047** (.0974,.1124)
MEAD c+p+liwc(psyc.)	.0626 (.0568,.0686)	.0931 (.0866,.0996)
MEAD c+p+liwc(pers.)	.0665 (.0594,.0736)	.0991 (.0911,.1069)
DUC2003M		
default MEAD	.0818 (.0701,.0935)	.1104 (.0975,.1228)
random	.0449 (.0371,.0530)	.0751 (.0680,.0823)
lead-based	.0681 (.0592,.0772)	.1000 (.0886,.1119)
MEAD c+p	.0778 (.0643,.0911)	.1106 (.0971,.1236)
MEAD c+p+liwc(all)	.0812 (.0686,.0933)	.1099 (.0986,.1214)
MEAD c+p+liwc(ling.)	**.0876** (.0745,.1015)	**.1206** (.1071,.1337)
MEAD c+p+liwc(psyc.)	.0791 (.0661,.0914)	.1082 (.0947,.1214)
MEAD c+p+liwc(pers.)	.0770 (.0645,.0894)	.1062 (.0941,.1178)
DUC2004M		
default MEAD	.0810 (.0722,.0899)	.1105 (.1021,.1197)
random	.0439 (.0378,.0500)	.0800 (.0728,.0872)
lead-based	.0762 (.0692,.0827)	.1069 (.0991,.1141)
MEAD c+p	.07745 (.0685,.0866)	.1086 (.0995,.1178)
MEAD c+p+liwc(all)	.0813 (.0728,.0896)	.1092 (.1007,.1174)
MEAD c+p+liwc(ling.)	**.0825** (.0733,.0909)	**.1114** (.1021,.1196)
MEAD c+p+liwc(psyc.)	.0753 (.0643,.0848)	.1044 (.0925,.1145)
MEAD c+p+liwc(pers.)	.0769 (.0678,.0854)	.1074 (.0971,.1172)

To further substantiate our claim, Figure 2 shows the average improvement of MEAD c+p+liwc(ling.) vs the baseline (in %) for ten groups of summaries of varying degree of difficulty. MEAD c+p+liwc(ling.) is highly effective for difficult cases but it is harming for easier cases. The outcome of this analysis seems to suggest that some documents or clusters can be effectively summarised using basic position and centroid features, while other documents or clusters require more evolved technology (for example, based on linguistic cues). By selectively applying one or another summariser, we would obtain a highly efficient summarisation strategy. In practice, to do so, we would need to predict the difficulty of every summarisation task (i.e. predict the ROUGE-2 score of the baseline). This is an interesting and novel line of future work.

4. DISCUSSION

Table 4 reports the sentence feature weights of the MEAD c+p+liwc(ling.) summariser. Some trends are consistent for both single-document and multi-document summarisation: the summariser gives preferences to sentences that i) have quantifiers, prepositions, conjunctions, impersonal pronouns, and ii) lack personal pronouns, 1st person plural, and adverbs. This fits well with some findings in the area of Psy-

[5] We repeated this analysis with other baseline summarisers and the conclusions remained the same.

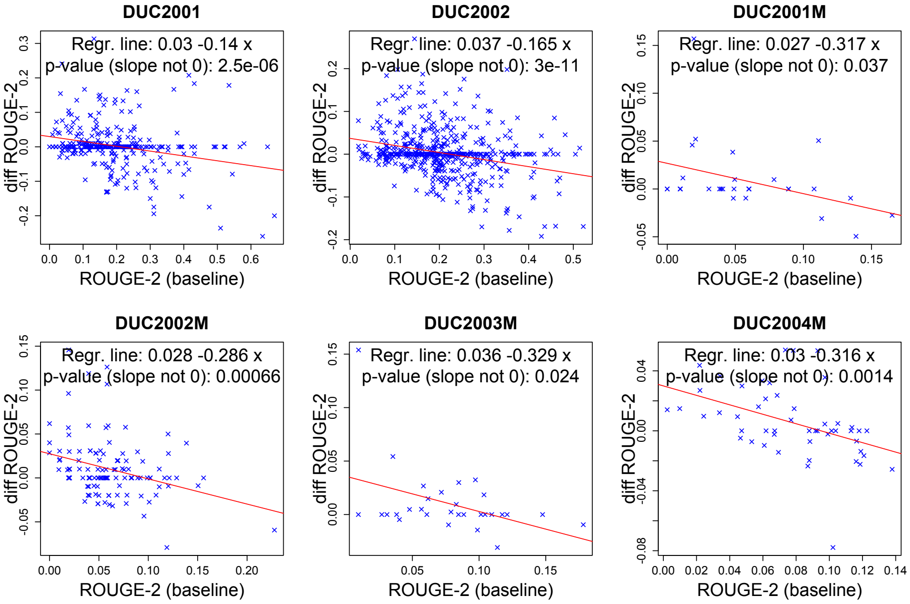

Figure 1: Regression analysis of the effect of the MEAD c+p+liwc(ling.) summariser. Each point in the plots represents a document (single-doc summarisation) or cluster (multi doc summarisation). The X coordinate is the ROUGE-2 score of the summary produced by a baseline summarizer (MEAD c+p tuned). The Y coordinate is the difference between the ROUGE-2 score of the summary produced by MEAD c+p+liwc(ling.) and the ROUGE-2 score of the baseline summary. In all cases, the negative association between X and Y is statistically significant. The MEAD c+p+liwc(ling.) summarizer works well with the summaries at the left end of the plot (difficult summarisation tasks) but it tends to deteriorate at the right end of the plot (easier summarisation tasks).

	Single-Doc	**Multi-Doc**
Centroid	0.236	0.666
Position	1.0	0.404
LIWC		
Total pronouns	-1.0	-0.924
Personal pronouns	-0.388	-1.0
1st pers singular	0.577	0.377
1st pers plural	-0.458	-0.887
2nd person	-0.878	0.993
3rd pers singular	1.0	-0.277
3rd pers plural	-0.302	-0.492
Impersonal pronouns	1.0	0.880
Articles	-0.397	1.0
Common verbs	-0.925	0.117
Auxiliary verbs	-0.079	-1.0
Past tense	0.102	-0.988
Present tense	-0.496	0.477
Future tense	0.983	-1.0
Adverbs	-1.0	-1.0
Prepositions	1.0	1.0
Conjunctions	1.0	0.879
Negations	-0.070	0.974
Quantifiers	1.0	1.0
Numbers	1.0	-0.690
Swear words	-1.0	1.0

Table 4: Weight of each sentence feature in the MEAD c+p+liwc(ling.) classifier. The weights are those fixed by PSO with the respective training collections.

chology of Natural Language about writing and analytical thinking. A higher use of conjunctions (for example, *but*, or *if*), prepositions (for example, *with*, or *over*), and quantifiers (for example, *some*, or *many*) is known to be associated with analytical thinking [18]. Analytical sentences need to make distinctions between ideas and, to do so, it is necessary to use these linguistic constructs. Our summariser gives more weight to sentences with conjunctions, prepositions, and quantifiers (very high weights to these three categories in both single-document and multi-document summarisation). This suggests that it is trying to promote analytical excerpts, which are arguably core elements of the text. The summariser's weights also give preference to sentences with more self-references and fewer verbs (high 1st person singular weight and negative common/auxiliary verbs weights). The use of first-person singular pronouns has been associated with people writing about real experiences [18]. Additionally, real stories, when compared to imaginary or fabricated stories, have fewer verbs (particularly, fewer auxiliary verbs that express discrepancy such as *would*, *could*, or *should*). A sentence with both stylistic elements (more "I-words" and fewer verbs) has more chance to be selected as a summary sentence. Again, this makes sense because the summariser is trying to extract document passages that describe real stories or events.

This analysis suggests that driving the summarisers with LIWC features has implicitly fomented analytical extracts and extracts about real experiences. From a language analysis perspective, this evidence reveals some unknown linguistic characteristics of salient sentences within news articles. Relating the psychometric profile of salient sentences

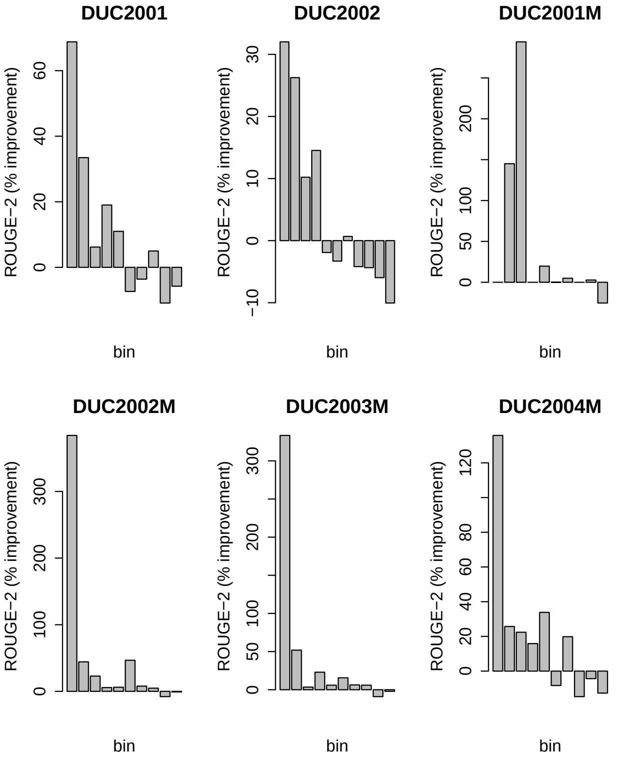

Figure 2: Percentage of improvement of the MEAD c+p+liwc(ling.) summaries from the baseline summaries (MEAD c+p). The summaries have been grouped into bins based on the baseline's performance (10% of the summaries in each bin). The left-most bin contains the 10% of the summaries having the lowest ROUGE-2 score (baseline), while the right-most bin contains the 10% of the summaries having the highest ROUGE-2 score (baseline). The Y axis represents the mean improvement of MEAD c+p+liwc(ling.) with respect to MEAD c+p (across all summaries in the bin). The MEAD c+p+liwc(ling.) summariser yields subtantial improvements for bins whose summaries have low ROUGE-2 performance (left-end of the plots).

to the discoveries in the area of Psychology of Natural Language has been insightful. In the future, we will further pursue this interdisciplinary analysis (for example, by extending this study to other types of text).

5. RELATED WORK

Two broad approaches to Automatic Text Summarisation have been identified in the literature [13]. *Shallow* approaches do not go beyond a syntactic level of representation and confine themselves to extracting salient parts of the text based on statistical, semantic –at word level–, or syntactic features. *Deeper* approaches employ natural language generation and semantic or discourse level representations. In Extractive Summarisation, a large majority of approaches follow the classic framework proposed by Edmundson [6], which is based on sentence scoring by weighted combination of several predefined features. Some well-known summarisation systems, such as MEAD [20] or SUMMARIST [8], follow this framework and incorporate new features or apply more sophisticated weighting methods. For instance, MEAD includes a re-ranking stage that removes redundant sentences following the Maximal Marginal Redundancy (MMR) principle [1].

To the best of our knowledge, our paper is the first attempt to inject psycholinguistic features into a state-of-the-art summariser. Psychological aspects of natural language use have been studied for a broad range of applications [19], e.g. to support analysis of emotions, personality, demography, health, deception, and other social or contextual variables. Word counting methods have helped to discover distinctive patterns of how people communicate in different situations. For instance, LIWC dimensions have been effectively used in opinion spam detection [15], sexual predator identification [16], or author gender identification [3]. A LIWC-based analysis was also recently conducted for predicting academic performance from students' written self-introductions [21].

Regarding evolutionary methods for summarisation, we have to refer to the recent work done by Kumar et al. [11]. In this paper, the authors address the multi-document summarisation task with Genetic Algorithms (GA). In a first step, they identify relations among documents; next, they score sentences from different documents taking into account the cross-document relations. Experiments against the DUC2002 dataset report improvements with respect to traditional cluster-based approaches. More recently, Khan et al. [10] also presented a method for multi-document summarisation that employs GA to weight features resulting from a semantic role labelling process. Again with the DUC2002 corpus, they also obtained improvements over existing summarisation methods.

Attacking Text Summarisation with psycholinguistic features is a novel and interdisciplinary way of approaching the problem. We expect that our current results stimulate discussion on this intriguing topic.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have provided preliminary empirical evidence on the effect of psycholinguistic features in Automatic Text Summarisation. Inspired by advances in the Social Sciences, we defined a novel set of features –related to psychological dimensions– and injected them into a state-of-the-art summarisation system. The resulting summarisation methods are slightly superior to existing summarisers. The improvements are modest but we believe that there is room for further enhancement. For example, by applying feature selection to individually extract LIWC features from every subset of LIWC dimensions.

We found that the summariser that includes linguistic LIWC dimensions is the best performing summariser. There are interesting connections between the occurrence of certain linguistic dimensions –for example, pronouns– and types of writing and thinking. We hope that our results serve as a basis to foster the discussion on how linguistic and psychological dimensions relate to sentence salience.

Another interesting finding is that our novel summarisation approaches are better suited for hard summarisation cases. This suggests that we could selectively apply standard summarisation methods or more advanced summarisers depending on the estimated difficulty of the summarisation task. To do so, we need to estimate the difficulty of summarising a given document or cluster. This challenge, which has clear connections with query difficulty estimation for Information Retrieval systems [2], will be subject to further research.

The test collections that we have worked with are essentially news datasets. In the future, we will also study how psycholinguistic features behave with other types of texts.

Acknowledgments

This work was supported by the "Ministerio de Economía y Competitividad" of the Goverment of Spain and FEDER Funds under the research projects TIN2015-64282-R.

7. REFERENCES

[1] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[2] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Morgan and Claypool Publishers, 2010.

[3] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[4] J. M. Chenlo, J. Parapar, D. E. Losada, and J. Santos. Finding a needle in the blogosphere: An information fusion approach for blog distillation search. *Information Fusion*, 23:58–68, 2015.

[5] R. C. Eberhart and Y. Shi. Comparing inertia weights and constriction factors in particle swarm optimization. In *Proceedings of the 2000 Congress on Evolutionary Computation*, volume 1, pages 84–88, 2000.

[6] H. P. Edmundson. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 2(16):264–285, 1969.

[7] D. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[8] E. Hovy and C.-Y. Lin. Automated text summarization and the SUMMARIST system. In *Proceedings of a Workshop on Held at Baltimore, Maryland*, TIPSTER '98, pages 197–214, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[9] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of IEEE International Conference on Neural Networks*, pages 1942–1948. Piscataway, NJ, 1995.

[10] A. Khan, N. Salim, and Y. Jaya Kumar. A framework for multi-document abstractive summarization based on semantic role labelling. *Appl. Soft Comput.*, 30(C):737–747, May 2015.

[11] Y. J. Kumar, N. Salimb, A. Abuobiedac, and A. T. Albaham. Multi document summarization based on news components using fuzzy cross-document relations. *Appl. Soft Comput.*, 21(C):265–279, Aug. 2014.

[12] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[13] I. Mani. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

[14] A. Nenkova and K. McKeown. A survey of text summarization techniques. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 43–76. Springer, 2012.

[15] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] J. Parapar, D. E. Losada, and A. Barreiro. A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs. In *PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse, at Conference and Labs of the Evaluation Forum CLEF*, Rome, Italy, 2012.

[17] J. Parapar, M. Vidal, and J. Santos. Finding the Best Parameter Setting: Particle Swarm Optimisation. In *2nd Spanish Conference on Information Retrieval (CERI 2012)*, pages 49–60, Madrid, Spain, 2012. Springer Verlag.

[18] J. W. Pennebaker. *The secret life of pronouns: what our words say about us*. Bloomsbury Press, New York, 2011.

[19] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577, 2003.

[20] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD – A platform for multidocument multilingual text summarization. In *Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

[21] R. L. Robinson, R. Navea, and W. Ickes. Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, 32(4):469–479, 2013.

[22] J. Steinberger and K. Jezek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2009.

[23] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.