Working paper

The Latent Semantic Scaling

Automated dictionary making technique for document scaling

Kohei Watanabe
06/06/2015

In computer-assisted text analysis (CTA), high portability is required to dictionaries, because they are usually repurposed with minor or no modification in different research settings. To achieve such portability, dictionaries have to be externally valid across different types of documents. The General Inquirer dictionary compiled by Stone (1966) is one of the highly portable CTA dictionaries, but dictionary created by automated techniques tend to be limited in portability. Wordscore (Benoit and Laver 2003) is a supervised method based on expert coding, and Wordfish (Slapin and Proksch 2008) is an unsupervised method using the maximum likelihood estimator. Although those two techniques proved their ability to scale party manifestos on left-right scale very close to human coding, they are generated from the same type texts that they are aiming to analyze, and highly specific to the type of documents used in training.

The technique developed by the author is mean to overcome the drawback of automated text analysis techniques by creating dictionaries from a large corpus of more general documents. It is call the *latent semantic scaling* (LSS), and based on a technique known as the latent sematic analysis (LSA), which was originally presented by Deerwester and others (1990) as part of their information retrieval system to overcome problems caused by synonyms and polysemy. The LSS has an ability to exploit information contained in a larger corpus to assign continuous scores to words for document scaling, while maintaining specificity of the dimensions that the scores represent. Since the dictionaries created by the LSS contain a large number of words extracted from large corpora, they are expected to have an ability to scale different types of documents on specific dimensions.

1 Algorithm

The LSS requires two sets of predefined words as human inputs. The first is words that specify concepts of interest (e.g. economy or immigration), and the second is words that define the dimension of scaling (e.g. left-right or positive-negative). The first set of words ('target words') is used to identify words related to the target concepts, and the second set of words ('seed words') is used to determine scores assigned to the relevant words.

Correspondingly, the LSS algorithm is made up of two steps. In the first step, entry words for a dictionary are chosen based on their association to target words. Those entry words are then scored by their distances to seed words in a semantic space reduced by the LSA. This two-step approach is analogue to the way early computerized content analysis dictionaries were created (see Martindale 1975; Pennebaker and Francis 1996; Stone et al. 1966).

For the example in this section, a corpus of UK and Irish news stories published between 1996 and 1997 on economy was constructed using Nexis database[1]. This corpus is consists of 55,263 news articles and the total number of words is 37 million.

1.1 Entry word selection

Entry words can be selected by using collocation measures, such as the point-wise mutual information (PMI) and the log-likelihood (LL). To create two-by-two contingency tables for the association measure, words in the corpus are distinguished into the local and foreign contexts. The local context only contains words appear within collocation windows from the seed words, and all other words belong to the foreign context. Collocation is often calculated using bigrams, but separation of words into two large chunks offers more accurate estimation since the system does not double count words when seed words appear more than once within a collocation window. The final result is not very sensitive to the size of collocation windows and it can range from 5 to 10, and entry words are typically

[1] Query for Nexis database was '(brit! OR UK OR irish OR ireland) AND econom! AND length(>100)' and sources were 'Irish Publications' and 'UK Publications'. Time window for the search was from 1996-05-01 to 1997-04-30.

modifier of the target words or words the target words modify, but not limited to those. Table 1 shows entry words extracted by the LL with window size of 5 using a wildcard expression 'econom*'.

Table 1: Entry words selected by econom*

Rank	Collocates	LL
1	growth	9969.0
2	social	5219.6
3	recovery	2264.4
4	political	2237.9
5	chief	1919.3
6	data	1518.5
7	inflation	1494.2
8	cycle	1462.6
9	rates	1346.8
10	policy	1338.0
11	activity	1336.9
12	growing	1083.8
13	indicators	1061.7
14	unemployment	1047.6
15	performance	1039.3
16	scale	1007.4
17	policies	959.9
18	strong	911.6
19	tiger	834.7
20	statistics	827.2
21	over-heating	823.9
22	strength	801.7
23	low-inflation	753.9
24	regeneration	752.5
25	boom	750.1
26	stability	749.9
27	miracle	735.1
28	global	733.8
29	inflationary	710.3
30	prosperity	697.6

Collocation analysis yields a large number of words, but we only adopt top 10% of collocates for entry words in English language, because only words that are very strongly associated with target concept are of our interests. The total number of entry words is 1,223 in this example.

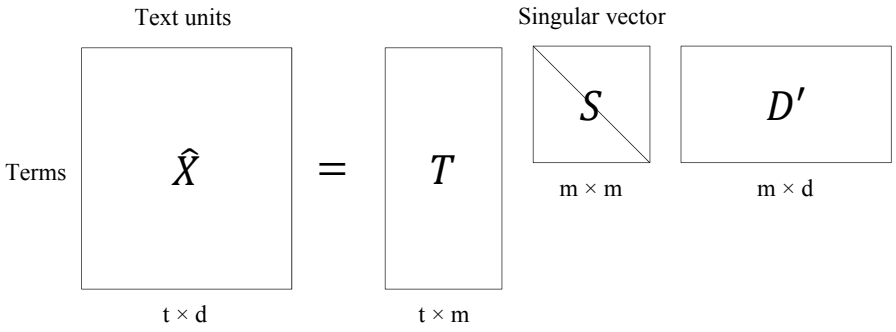
1.2 Word scoring

Scores of entry words are calculated based on their semantic distance to seed words. Seed words can be any set words, but they have to be in two subsets that represent both ends of linear dimension as pairs. For example, general English positive-negative seeds created by Turney and Littman (2003) are {good, nice, excellent, positive, fortunate, correct, superior} for the positive end and {bad, nasty, poor, negative, unfortunate, wrong, inferior} for the negative end.

The relationship between the two subsets is then translated into seed scores. Words in one subset are given scores totaling +1, and those in the other subset are −1 in total. For the positive-negative seeds, scores become inverses of seven as following: {good: +0.142, nice: +0.142, excellent: +0.142, positive: +0.142, fortunate: +0.142, correct: +0.142, superior: +0.142, bad: −0.142, nasty: −0.142, poor: −0.142, negative: −0.142, unfortunate: −0.142, wrong: −0.142, inferior: −0.142}. Although the numbers of seed words are equal in this example, subsets can be in different sizes.

The documents in the corpus are split into smaller units (sentences or paragraphs), and a term-unit matrix is created. Yet, since the original matrix[2] is very large and contains unessential information, the singular vector decomposition (SVD) is performed to reduce the size. The original matrix X is reduced to $\hat{X}$ by SVD:

$$X \approx \hat{X} = TSD'$$



The reduced term-unit matrix $\hat{X}$ now represents a semantic space in which distances between the seed words and entry words are measured. Term-term similarity in this matrix is $\hat{X}\hat{X}' = TS^2T'$, but, we can also obtain term-term similarity by taking the dot-products between two rows vectors in TS while avoiding creation of a large dense matrix $\hat{X}$ (Deerwester et al. 1990). Yet, unlike the standard the formula, the LSS utilize cosine similarity of row vectors from TS as term-term similarity measure to

[2] Turney and Littman (2003) note that the original matrix are usually Tf-idf transformed in the LSA but it deteriorated the performance of the LSS, so original frequency count was retained.

avoid dominance of highly frequent seed words. The valence score v_i for an entry word w_i is the mean cosine similarity to seed words weighted by the seed scores:

$$v_i = \frac{1}{n}\sum_{j}^{n} \mathrm{COS}(w_i, s_j) \cdot e_j$$

where $\mathrm{COS}(w_i, s_j)$ is cosine similarity between row vectors of TS for a word w_i and a seed word s_j, and e_j is seed score of s_j. Note that since cosine similarity ranges between -1 to 1, negative similarity to a negative seed word yields a positive score, and words received highest scores are those close to positive seeds and distant from negative seeds. This scoring method consequently restricts the scale to the positive-negative dimension and promotes words that are in the semantic space between the negative and positive seed words.

Figure 1 illustrates the scoring method of the LSS. The horizontal line is the positive-negative dimension defined by the seed words, and the arc represents positions in the semantics space that are in equal distance to negative seeds. Words w_1 and w_2 are in the same distance from the negative seeds, but w_2 is off the horizontal line. As two concentric circles around the positive seeds shows, due to the deviation from the straight line, the distance from positive seeds to w_2 are greater than to w_1, and thus scores for w_2 becomes smaller and less significant than w_1.

Figure 1: Concept of LSS scoring method

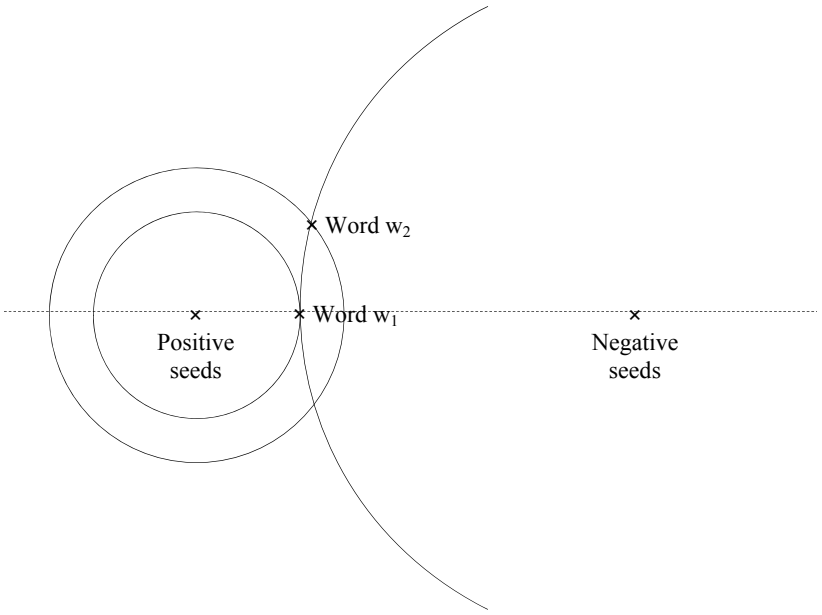


Table 6 shows positive and negative words on economy. The positive and negative words are among the 1,184 words in the dictionary, but only most significant words are selected for presentation. Scores are rescaled by multiplying by a constant that makes the highest score equal to 100 for human readability, and range from -33.75 to 100. We do not find any of the entry words presented in Table 2, because their valence scores are not among the extreme values despite their strong association with the target word, 'economy'.

Table 2: positive and negative economic words

Rank	Positive words		Negative words	
1	good	100.00	poor	-33.75
2	achieving	71.68	damage	-31.46
3	prospects	71.01	politics	-27.49
4	role	68.77	news	-26.39
5	successful	67.84	yen	-24.78
6	strategic	64.17	devalue	-24.08
7	reasonably	64.04	voodoo	-23.42
8	shape	63.95	countries	-21.45
9	rewards	62.77	currency	-21.07
10	potential	62.49	dollar	-20.99
11	ensure	61.70	soaring	-19.80
12	sufficiently	60.57	suffer	-19.15
13	skills	60.50	devaluation	-19.01
14	dynamic	60.23	caused	-18.17

15	opportunities	58.61	blow	-17.64
16	indicator	58.54	health	-17.54
17	sustainable	58.29	new-right	-17.09
18	continue	58.20	sterling	-15.15
19	assessment	57.82	warn	-15.12
20	essential	57.72	fluctuations	-14.72
21	underpin	57.61	rising	-14.32
22	leading	57.07	catastrophic	-13.79
23	consistent	56.97	hinder	-13.37
24	objectives	56.29	deprivation	-13.22
25	sustainability	55.56	low	-13.21
26	importance	55.16	crippled	-13.01
27	achievement	53.24	pound	-12.75
28	believes	53.06	collapse	-12.03
29	feasible	52.24	blockade	-11.32
30	robust	51.50	unrest	-11.19

2 Experiments

In the first experiment, I will first attempt to measure economic policy positions of the UK and Irish parties in their manifestos for 1997 elections by a LSS dictionary created from the UK and Irish economic news corpus. This analysis is meant to be a replication of the Wordscore paper (Benoit and Laver 2003). Then, the same procedure is repeated using UK news corpora from six different time periods to test external validity of the method. Finally, the dictionary created from the UK and Irish corpus is used to scale the UK manifestos from 1995 to 2010 to demonstrate its portability. The left-right position scores for this experiment was adopted from more recent paper (Benoit et al. 2014), although these scores are not exactly the same as those presented in the Wordscore paper.

The second experiment focus on more specific dimension in party manifestos, and positive-negative attitude toward immigration will be measured using the LSS. Positive-negative dictionary on immigration will be created from a corpus constructed from the UK newspaper articles published during the one year period before the 2010 general election. The data for this experiment was produced by Kenneth Benoit by using Amazon Mechanical Turk. This experiment is meant to illuminate the advantage of the LSS in measuring a very specific dimension in documents, and to serve as a substitute for social policy scaling in the Wordscore paper.

In the following two experiments, identical methods (the same algorithm and parameters) were used to avoid over-fitting. Entry words were top 10% of collocations of target words selected by the Log-likelihood measure with a window size of 10. In the SVD, the number columns of reduced term-unit matrices was set to 300, the known optimal size for synonym extraction (Landauer and Dutnais 1997). More practical issuers will be discussed in the latter section.

2.1 Left-right position on economic policy

In order to a create left-right position dictionary by the LSS, a set of economic position left-right position seed words was created by the author. It is comprised of {deficit, austerity, unstable, recession, inflation, currency, workforce} for the right and {poor, poverty, free, benefits, prices, money, workers} for the left. Those words were selected manually while reading the UK and Irish manifestos, but not necessarily appear in those documents. The rightist words contain words that are often used when macro-economic issues are mentioned, whereas the leftist words are used when micro or individual level economic problems are addressed. The top-30 words in the left-right position dictionary are presented in Table 3.

The result of the computer coding by the left-right dictionary is shown in Figure 2. In the chart, UK and Irish parties are found along the 45-degree line, indication strong correspondence between scores assigned by experts and the LSS dictionary. The overall agreement between expert and LSS scores measured by Pearson's correlation coefficient is r=0.946. The statistical significance of the correlation coefficient was tested by empirical distribution obtained by simulation (Figure 3). To obtain the empirical distribution, ten thousand dictionaries from randomly chosen sets of seed words (7 words for each left and right) was created and computer coding was performed in the same way. The density of correlation coefficients peaks at around -0.5 and 0.5, but it drops sharply toward the extremes. As a result, the critical value for the 95% and the 99% confidence levels (two-tail) are respectively r=0.806

and 0.902. Therefore, we can confirm that the correspondence between the experts and LSS scores is statistically significant.

Table 3: Rightist and leftist words on economy

Rank	Rightist words		Leftist words	
1	deficit	100.00	benefits	-75.39
2	medium-term	99.43	poor	-72.25
3	deficits	95.05	benefit	-56.32
4	exchange-rate	95.00	health	-48.06
5	devaluation	89.70	migrant	-46.18
6	central-bank	88.89	dependency	-43.40
7	finance-minister	86.28	improve	-41.54
8	austerity	84.62	inequality	-41.27
9	gross-domestic-product	83.43	deprivation	-39.37
10	exchange-rate-mechanism	83.42	equality	-39.21
11	outlook	83.05	disadvantaged	-37.84
12	budget	82.65	welfare	-37.25
13	forecast	77.93	society	-35.77
14	strength	76.58	socially	-35.61
15	recession	76.11	exclusion	-34.86
16	monetary-union	75.96	individualism	-33.95
17	prudent	75.52	cohesion	-33.40
18	deflationary	74.75	wealth	-32.34
19	projections	74.68	social	-31.65
20	weakening	74.60	fabric	-31.02
21	sterling	73.76	well-being	-30.24
22	stoking	73.03	inequity	-30.08
23	pound	71.76	performance	-30.05
24	inflation	71.39	communities	-28.50
25	tight	70.18	trickle-down	-28.39
26	devalue	70.05	reap	-27.12
27	forecasters	69.94	incomes	-27.05
28	forecasts	69.58	welfare-state	-26.77
29	underlying	67.75	freedoms	-26.57
30	expansionary	67.14	wage	-26.51

Figure 2: UK and Irish manifestos coded by dictionary created from 1997 UK and Irish news corpus

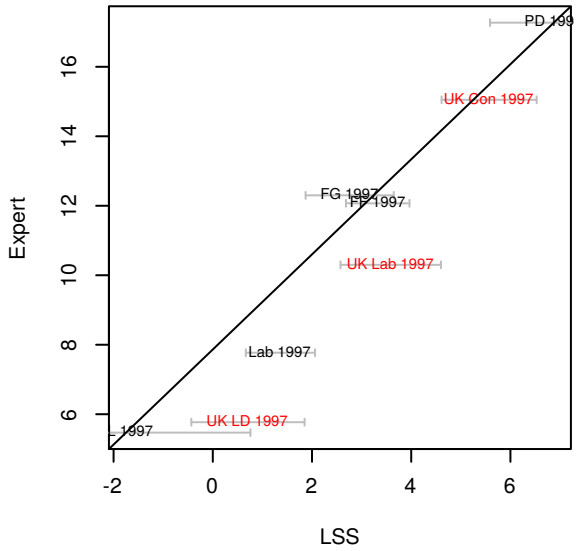
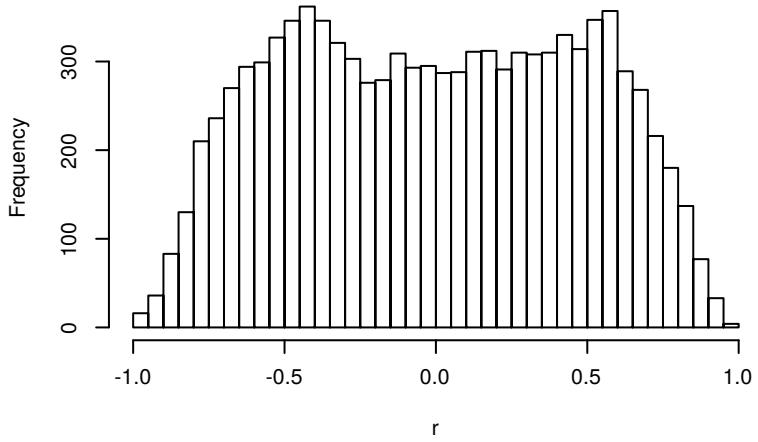


Figure 3: Distribution of correlation coefficient by random left-right position dictionaries



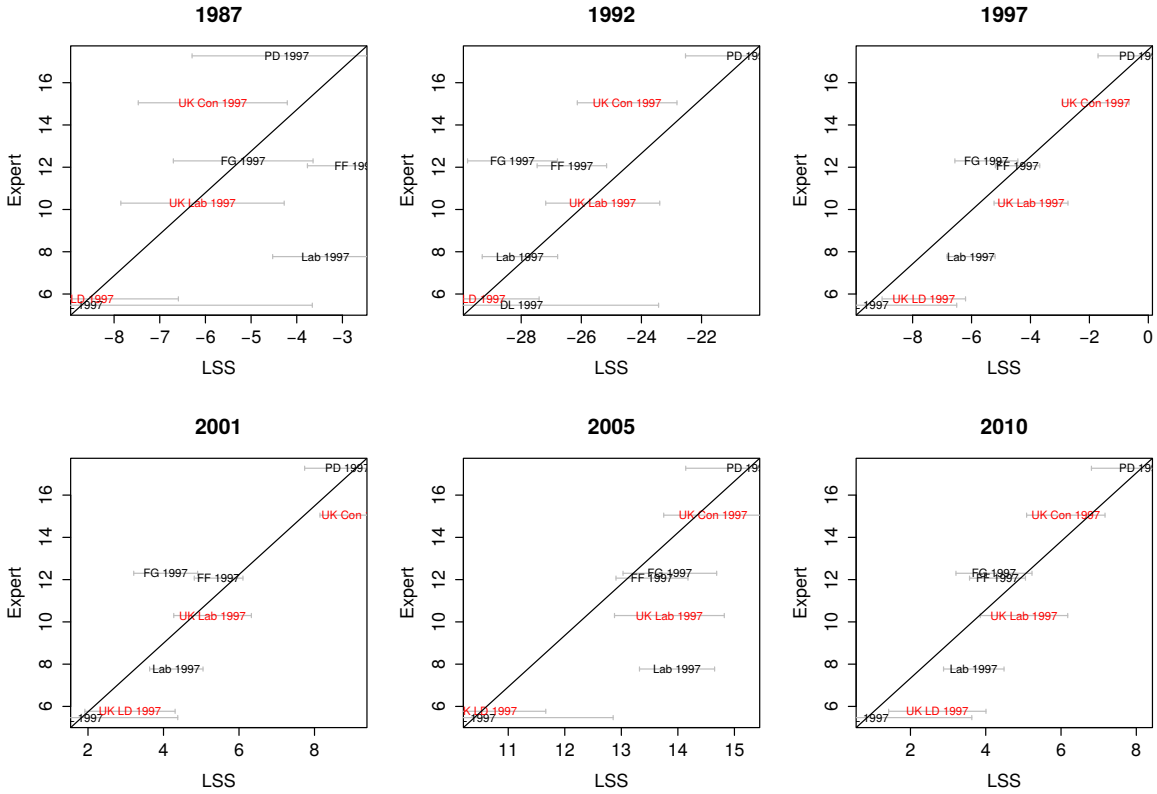
Additionally, a robustness test was also performed by randomly taking 25%, 50%, 75%, and 100% of document from the corpus with replacement. When only 25% of news articles are subsampled, 49 out of 100 dictionaries achieved correlation higher than the critical value (p=0.01); when it was 50%, 87 dictionaries achieved beyond the level; when the size of the corpus is increased to 75%, 99 cases

exceeded the threshold. If the size of the corpus is the same as the original (100%), none of the cases was below the threshold. This result seems to be indicating the robustness of the LSS when the corpus size is over 18.5 million words, half of the original corpus size.

In order to test external validity, six more left-right dictionaries were then created from different corpora consisting of UK newspaper articles[3] on economy published during the one-year periods prior to general elections in 1987, 1992, 1997, 2001, 2005 and 2010. The correspondence between experts and LSS scores are presented in Figure 4 and their correlation coefficients were respectively 0.568, 0.840, 0.944, 0.911, 0.864 and 0.928. The correlation coefficient for the 1987 corpus (r=0.568) is even below the 95% significance level (r=0.806), but all others are above this level; they become even stronger after 1997 and exceeds 99% confidence level except for 2005 corpus (r=0.864). The inability of the LSS to produce similar results from 1987 corpus can be explained by its small size. The corpus only contains 11,486 (8.6 million words) stories, while others contain respectively 24,111 (17.3 million), 28,162 (21.4 million), 32,251 (24.0 million), 26,223 (20.1 million), and 43,254 stories (33.8 million). The problem of the 2005 corpus is not very clear but the dictionaries created from the corpus always underperformed. I suspect that the news stories retrieved by database is limited for copyright reasons or affected by some technical failure in Nexis, and eventually incomplete, considering the fact that number of news stories (26,223) is fewer than the 1997 corpus (28,162) despite the general tendency that number of stories increases over time.

[3] When those newspaper stories were downloaded from Nexis database, publications were limited to those existed in 1997, because the number of publications available in the database is much greater after 2000.

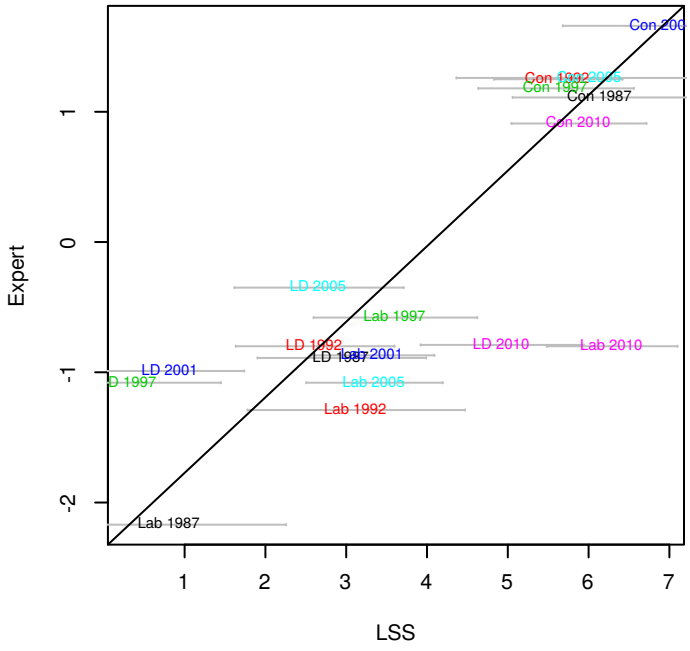
Figure 4: Dictionaries created from UK news corpus from different time periods (1978-2010)



As a test of portability, the dictionary created from the UK and Irish newspapers in 1997 was used to scale UK manifestos from 1987 to 2010. The economic policy positions scores were adopted from expert coding presented in a paper by Benoit and others (2014). In Figure 5, we can observe strong correspondence between expert and LSS scores (r=0.797) and clear separation between the Conservative (Con) and the Labour (Lab) and the Liberal Democratic (LD) by the LSS, although not all of them are positioned along with the diagonal line. Large deviation from the diagonal line, however, can be found in the LD 1997 and the LD 2001. The deviation of the LD 1997 is due to the difference in score assigned to the document from that in the Wordscore paper. In the Wordscore paper, the distance between the LD 1997 and the Lab 1997 are approximately equal to the Lab 1997 and the Con 1997 (Figure 3), but they are much close here. There is no evidence, but the same explanation applies to the LD 2001. Its score could have been lower if exactly the same coding scheme had been employed. The LD 2010 and Lab

2010 are also showing large deviation, but it is arguably reflecting the substantive change in their economic policy after the 2008 economic crisis. Namely, after the crisis, the political leaders to address national and international economic problems, and even leftists parties' economic policies became hardly distinguishable from the Conservative from the perspective of the 1990s politics.

Figure 5: UK manifestos coded by dictionary created from 1997 Irish and UK news corpus



2.2 Positive-negative attitude toward immigration

Creation of dictionary that measure attitude toward immigration does not require new seed words, but we can simply use the above presented general English positive-negative seed words (Turney and Littman 2003). Entry words for this immigration dictionary were extracted using a wildcard expression 'immigra*' and 'migra*'. The corpus for this dictionary was created by downloading all the UK news

stories on immigration published between May, 1, 2009 and April, 30, 2010 from Nexis database[4]. The total number of stories in the corpus is 15,343 containing 11.6 million words in total. Simple keywords searches retrieved stories about animal migration and headache (migraine), but impact of those stores were limited.

Table 3 presents most positive and negative words in the immigration dictionary. We can notice that many of the words are intuitively positive or negative but some are neither positive nor negative by itself. For example, 'globalization' is usually a neutral word, but it is positive in the context of immigration because the word is often used when people discuss advantage or inevitability of immigration to the UK. Contrary, 'control' is a negative word for immigration since it is frequently used in an argument for suppressing the number of migrants to the UK. Words such as 'breed', 'species' and 'wildebeest' are present because Nexis database retrieves stories about animal migration, but they have minimal impact on coding since political documents rarely contain those terms.

Table 3: positive and negative words on immigration

Rank	Positive words		Negative words	
1	skills	100.00	xenophobia	-141.27
2	globalisation	88.24	control	-130.09
3	chauffeured	86.93	racist	-125.20
4	airport	86.68	stemming	-122.50
5	ranging	82.41	tide	-122.46
6	clearance	79.48	working-class	-115.53
7	status	78.40	negative	-113.76
8	agency	74.98	failure	-110.32
9	issues	72.15	problems	-106.95
10	breed	69.45	influx	-100.81
11	claimed	68.84	branded	-99.42
12	vehemently	68.60	caused	-96.82
13	skill	67.30	exploit	-94.11
14	test	65.91	first-generation	-90.78
15	attract	64.39	warned	-89.93
16	permanent	63.68	families	-88.51
17	legal	59.23	soaring	-86.53
18	melting-pot	57.34	ignored	-86.45
19	species	57.27	housed	-85.33
20	wildebeest	56.96	magnet	-84.47
21	overstaying	56.07	borders	-83.18
22	documents	55.90	newly-arrived	-83.12
23	routes	55.75	accused	-82.89
24	work	55.63	evicted	-82.02

[4] The search query was 'brit! AND (immigra! OR migra!) AND length(>100)' and source was 'UK newspapers'.

25	shambles	55.28	trickle	-81.42
26	breeding	53.65	rates	-79.42
27	bringing	53.24	fuelled	-78.34
28	employ	52.76	flooded	-76.69
29	passport	52.24	non-white	-76.48
30	official	51.88	lorries	-76.38

Human-coded data was created by Ken Benoit using Amazon Mechanical Turk (MT), in which participants were asked to judge whether sentences are about immigration and positive (+1), neutral, or negative (−1) on the issue. In creation of the MT scores, only sentences that judged as about immigration by more than two third of coders were selected, and positive-negative scores assigned to the sentences were averaged across coders. Those sentences were then combined to create manifestos on immigration, and their overall scores were calculated by taking average score of all the sentences in respective documents.

Correspondence between scores assigned by the human coders and the LSS is presented in Figure 6. Although immigration is only briefly mentioned in those documents, and thus the 95% confidence interval is wide, all the parties are placed along the diagonal line (r=0.915) except for the Green party. The significance of the correlation is tested by creating ten thousand dictionaries (Figure 7), and critical values for 95% and 99% in the two-tailed empirical distribution were r=0.773 and 0.859. The correlation between the human and LSS coding is way above the 99% confidence level.

Figure 6: UK manifestos on immigration selected by human coders

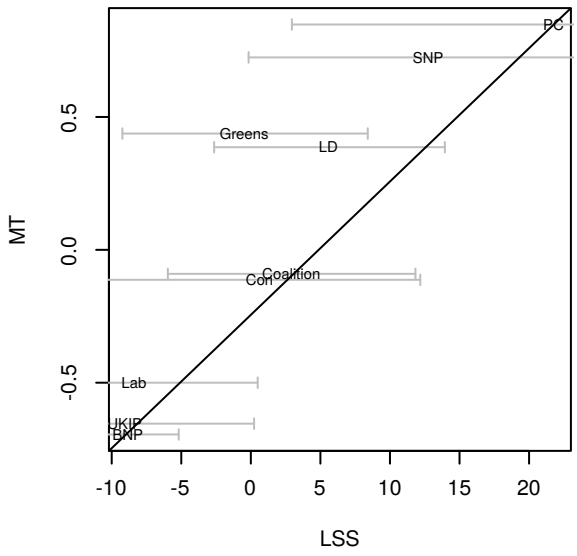
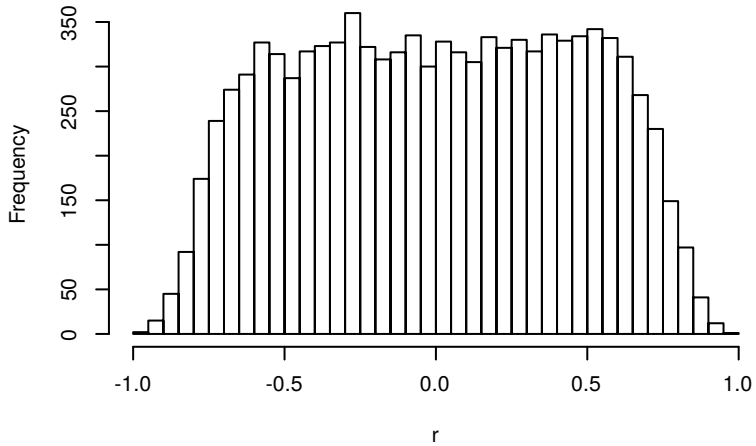
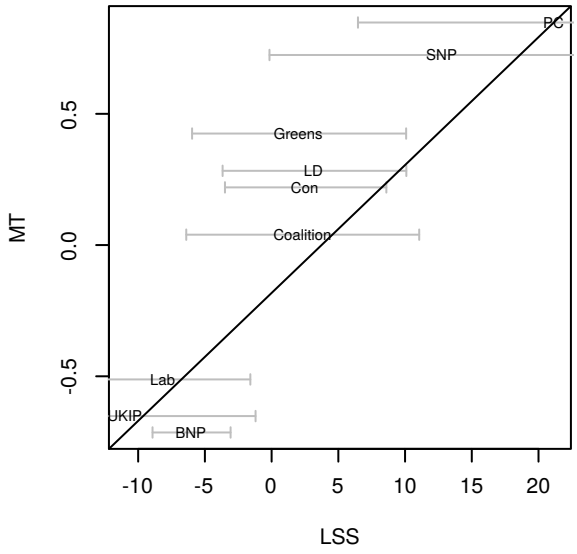


Figure 7: Distribution of correlation coefficient by random immigration dictionaries



However, in reality, selection of specific sentences by human coders is not always feasible. Hence, I extracted immigration sentences by simple keyword matching. The keywords used for this information retrieval (IR) approach were ‘immigra*’, ‘migra*’, ‘refugee*’, ‘asylum*’, ‘foreign*’. The result presented in Figure 8 is very similar to previous result, and we can observe strong correspondence (r=0.932) between the aggregated manual scores and the LSS scores corresponds.

Figure 8: UK manifestos on immigration selected by keywords



3 Practical issues of LSS

Use of computerized text analysis techniques in social scientific research is not only about algorithms, but also about corpus construction, unitization and tokenization. This section will discuss the practical issues of the LSS for its successful application, and they will be followed by brief notes on implementation of the system.

3.1 Corpus

The LSS require a corpus that contains a large number of documents on subjects of interests. As we have seen in the bootstrapping of UK and Irish economic news corpus, when the number of news stories are halved to 27,631 (approximately 11 million words), the number of dictionaries that achieve statistically significant correlation with manual coding at the 99% confidence level drops from 100 to 87; when only the quarter of the documents are used (13,815), only 49 dictionaries reached the level. The impact of small corpus size can be also seen in the difference in performance of dictionaries created from the UK corpora from 1987 and 1992. The numbers of documents in those corpora are respectively

11,486 and 24,111, and their performance measured by Pearson's correlation coefficient was 0.568 and 0.840.

Nevertheless, it seems that smaller corpora are sufficient when more common words are used as seeds. For the creation of the immigration dictionary, a corpus only contains 15,343 documents (11.6 million words) was used but the results was satisfactory achieving strong correspondence ($r=0.932$, $p<0.01$) with human coding.

In most of the research situations, online databases such as Nexis seem to be the best tool for construction of time and topic specific corpus, but search terms have to be selected with precaution. Most common problem in corpus construction by databases is contamination by irrelevant documents. It was animal and headache related stories in the case of our immigration dictionary, although their impact was limited. More dramatic and serious problem is a lack of homogeneity in a corpus. For example, search terms 'uk' and 'brit*' are equivalent words in our common sense, but they retrieve largely different set of documents on immigration, because 'uk' is used in news stories that are more positive or neutral on the topic[5]. If those two search terms are used in a single query, the corpus lacks homogeneity and dictionaries created by the LSS do not perform well.

Finally, it is noteworthy that duplicated sentences have only limited impact on the performance of the dictionaries. I initially expected that the sequences of words that are quoted or repurposed (typically, press releases and news wires) in many news stories bias the collocation association measure, but it wasn't really the case. Therefore, we do not need to exclude sentences that repeatedly appear in the corpus using complex mechanisms.

[5] Immigration dictionaries created from the two corpora showed that 'worker' and 'employing' appear with opposite valence signs: 'worker' is the second most positive word in UK corpus (+84.5) after 'checks' (+84.5), while it is negative in British corpus (−1.1); 'illegal' is scored −43.9 in British corpus, but it has almost neutral score (+0.9). A word 'airport' is the second most positive word (+98.8) the dictionary created from the British corpus, but it does not appear in the dictionary created from the UK corpus. Further, the standard deviations of the scores in the dictionary created from the British corpus ($\sigma=40.6$, $n=746$) is twice as great as that from the UK corpus ($\sigma=23.1$, $n=612$).

3.2 Seed words

One of the most important components of the LSS is seed words, which are exemplary supervision to the system in this computer assisted dictionary making technique. If appropriate seed words already exist, we can adopted them to create new dictionaries from different corpora, because the LSS can produce very similar results even from very different corpora (Figure 5), but unfortunately, there are only a two sets of seed words known to us (positive-negative and left-right).

Creation of new sets of seed words is a great contribution to our community and strongly encouraged, but it is not an easy task. Seed words should be a set of words that represent opposite ends of dimension of interest; each set have to have more multiple words, ideally more than five for each end, for stable outcomes; they have to be very specific[6] and inclusion of ambiguous words is destructive[7].

As rule of thumb, those who wish to create a new set of seed words manually should start from choosing a pair of two words that are representing opposite ends on the dimension of interest, create a dictionary and test them with human coded data; if the pair is deemed valid, those words can be retained and a next pair should be looked for; this process shall be repeated for many times. Choice of seed words are creative and exploratory process, but scanning of target documents and referring to general thesaurus can be useful. After discovering five or more pairs, they can be combined into a set of seed words.

However, there are two potential ways to create seed set with the help of computer programs. If there are documents know to be different on the dimension of ones interests, relative frequency of words can be used to extract seed words. For example, seed words extracted from the 1997 UK and Irish party manifestos based on comparison of word frequencies[8] between two groups of documents that are separated by the median expert scores produces very similar results as the manually selected seed words

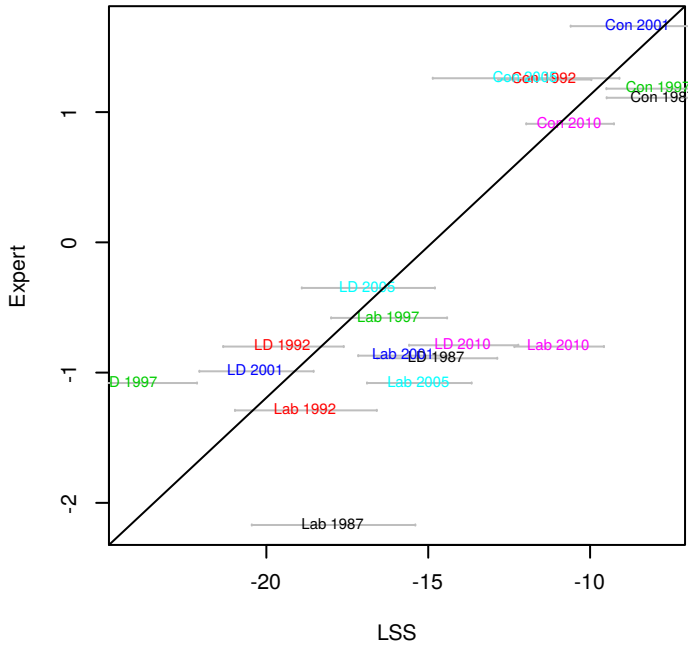
[6] For instance, in the left-right seed words, 'prices' were chosen over 'price', because we are interested in general living costs rather than price of individual goods.

[7] For example, when seed words are specified by wildcard expressions (e.g. 'benefit*' or 'price*') to increase the number of seed words, the performance of the dictionary usually decreases.

[8] In the word selection, word frequencies were measured by the quadratic point-wise mutual information (PMI3), and rightists and leftists party manifestos were used as target and reference documents; words with highest and lowest 50 PMI3 scores were chosen as seed words.

in the LSS. The dictionary created from word frequency-based seeds can scale documents correspondingly to the expert scores (r=0.88, p<0.05), and it can also produce an almost identical result when applied to the 1992-2010 UK manifestos (Figure 9). This approach is very close to Laver and Garry (Laver and Garry 2000), but the lexical diversity and thus generalizability of this dictionary is much greater than their dictionary thanks to the large corpus.

Figure 9: UK manifestos coded by dictionary created from seed words extracted from the 1997 Irish and UK manifestos



Another prospective way to extract seed words automatically is use of corpora comprised of documents known to us. For instance, if the general positive-negative seeds are used with Irish economy news corpus from 1997, the dictionary can scale left-right positions of the Irish party manifestos in a way that moderately corresponds to human coders (r=0.74). It occurs because Irish economy was booming in the 1990s under the liberal economic regime and thus words associated with rightist economic policy were positively mentioned in newspapers. This technique was not tested much in this paper, but it is worth exploring in the future.

3.3 Tokens and units

In computerized text analysis, word stemming is often used to reduce the number of token types, but stemming is not necessary in the LSS. The importance of retaining original tokens in the LSS is arguably because semantic distances are measured by similarly in contexts in which words appear, and inflected forms allow the system to distinguish between different contexts. Further, stemming prevent us from selecting very specific words and increases the harmful ambiguity of seed words.

In a larger corpus, there are usually large variations in spelling of words, especially of compound words such as 'melting-pot', 'newly-arrived', 'working-class', and 'non-white' in the immigration dictionary (Table 7). Since these words are relatively new, they tend to appear in different notations either intentionally or erroneously: with hyphen ('melting-pot'), without hyphen ('melting pot'), or concatenated ('meltingpot). For accurate estimation of words association and semantic distance, spelling of compound words have to be standardized in every corpus. This requires extensive pre-processing of corpora, in which all the variants are converted into hyphenated formats as simple strategy in the current system[9].

One of the most important findings in this research was the need of exclusion of named entities from corpora to yield good results: if named entities were not removed, the performance of the dictionary was way below the above presented level. This can be explained by the fact that news stories frequently mention to persons, organizations and places, and their names are fed into the LSA as contextual information. For removal of proper nouns, we need to add a named entity recognition mechanism into pre-processing of corpora. In the current system, proper names were recognized simply by frequency of capitalization in combination with a more complex multi-part names tokenization mechanism developed based on the technique presented by Blaheta and Johnson (2001).

[9] If hyphenated formats are less frequent than one in million tokens in the original texts, they are ignored.

Finally, term-unit matrices should not be constructed from original documents (e.g. news stories) as unit, but from sentences of the documents, because one whole document contains a diverse set of dimensions on a certain subject. Paragraph was a possible unit in the experiment, but the LSS performed when documents were reunitize into sentences. Unitization of larger corpora results in extremely large term-unit matrices, but these are manageable if they are stored as sparse matrices.

3.4 Implementation

Despite the simple algorithm, implementation of the LSS requires a number of subroutines for the intensive pre-processing. The current system is implement entirely in Python and the LSA is performed using the Gensim library (Řehůřek and others 2010). The flow of dictionary making in the current system is the following:

1. Unitization of texts
2. Tokenization
    a. Text cleaning
    b. Stop-words removal
    c. Compound words identification and standardization
    d. Named entity recognition and elimination
3. Entry word selection by collocation analysis
4. Word scoring by the LSA

The key issue in implementation of the LSS is the capability to treat large corpora that contain tens of million words. The current system running on Ubuntu Linux OS performed well for the pre-processing in terms of speed, but required approximately 8GB of RAM. Gensim library could be used for the large corpora thanks to its online algorithm that process texts piece-by-piece with a limited memory usage, although its power iteration parameter needed to be set to 10 for highly reliable outputs, where default is 2. As a result, the whole process of the LSS dictionary making took 20-25 minutes. This

may sound very long, but it is dramatically quick dictionary making, if we remember that manual dictionary making takes weeks or months by a team of researchers.

4 Conclusion

The LSS was able to replicate the expert coding of the 1997 UK and Irish manifestos, the data used for the Wordscore paper. The dictionary was also able to separate left-right positions of manifestos of three major UK parties from 1987 to 2005. This seems to be indicating the internal and external validity of the LSS. Such an external validity implies portability, capability of LSS dictionaries to be repurposed in other research projects, that has been the characteristics manually compiled CTA dictionaries.

The result produced by the immigration dictionary is indicating the ability of the LSS to measure very specific dimension in documents. In Bayesian supervised learning, creation of large dictionaries on a specific topic demands a sizable training data, but the LSS is free of such costly training, because it can extract relevant words from a corpus and scale them using seed words in the semantic space.

Since the LSS is corpus-based dictionary making technique, construction of large complete corpora is essential. As we have seen in the bootstrapping, the performance of a dictionary quickly deteriorates as the size of corpus shrinks; if a corpus does not contain complete set news stories, the LSS fails to produce valid dictionary. For corpus construction, online news databases are very useful tool to construct large time and subject specific corpora, although they are not always perfect as we have seen as in the UK economic news corpus in 2005.

Seed words creation is the key to the LSS, but discovery of new seed words is iterative process just like manual dictionary creation, and thus techniques to find seed words need to be developed further. Yet, one's time and energy invested into creation of seed words never be wasted, because they can be repurpose for different corpora and produce myriad of new dictionaries by other researchers.

Additionally, we can seek for a technique to making a dictionary from corpora whose contents are known to us without creating new seed words.

Finally, one of the most interesting features of the LSS is its strong specificity in word scoring. Seed words created for a corpus from a certain time period (e.g. 1997 UK corpus) retain the dimension even in the corpora that are very different in nature (e.g. 2010 UK corpus). Similarly, dictionaries created to scale documents from a certain time period (e.g. 1997 UK and Irish manifestos) do not distinguish between documents from very different tine period (e.g. 2010 UK manifestos). The specificity of the LSS dictionary confirms the high reliability of computerized content analysis, and its general advantage over manual content analysis.

Bibliography

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2014. "Crowd-Sourced Coding of Political Texts." https://files.nyu.edu/ml127/public/Forthcoming/Crowd%20sourced%20data%20coding.pdf.

Benoit, Kenneth, and Michael Laver. 2003. "Estimating Irish Party Policy Positions Using Computer Wordscoring: The 2002 Election – a Research Note." *Irish Political Studies* 18 (1): 97–107. doi:10.1080/07907180312331293249.

Blaheta, Don, and Mark Johnson. 2001. "Unsupervised Learning of Multi-Word Verbs." In *Proceeding of the Acl/Eacl 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, 54–60.

Deerwester, Scott C., Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *JASIS* 41 (6): 391–407.

Landauer, Thomas K., and Susan T. Dutnais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review*, 211–40.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44 (3): 619. doi:10.2307/2669268.

Martindale, Colin. 1975. *Romantic Progression : The Psychology of Literary History*. Washington, DC: Hemisphere Publishing ; New York ; London.

Pennebaker, James W., and Martha E. Francis. 1996. "Cognitive, Emotional, and Language Processes in Disclosure." *Cognition & Emotion* 10 (6): 601–26. doi:10.1080/026999396380079.

Řehůřek, Radim, and others. 2010. "Fast and Faster: A Comparison of Two Streamed Matrix Decomposition Algorithms." http://nlp.fi.muni.cz/~xrehurek/nips/rehurek_nips.pdf.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. doi:10.1111/j.1540-5907.2008.00338.x.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The M. I. T. Press.

Turney, Peter D., and Michael L. Littman. 2003. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association." *ACM Trans. Inf. Syst.* 21 (4): 315–46. doi:10.1145/944012.944013.