Exploring the Traits of Manual E-Mail Categorization Text Patterns

Eriks Sneiders, Gunnar Eriksson, and Alyaa Alfalahi

Stockholm University, Department of Computer and Systems Sciences,
Forum 100, SE-164 40, Kista, Sweden
{eriks,gerik,alyalfa}@dsv.su.se

Abstract. Automated e-mail answering with a standard answer is a text categorization task. Text categorization by matching manual text patterns to messages yields good performance if the text categories are specific. Given that manual text patterns embody informal human perception of important wording in a written inquiry, it is interesting to investigate more formal traits of this important wording, such as the amount of matching text, distance between matching words, n-grams, part-of-speech patterns, and vocabulary in the matching words. Understanding these features may help us better design text-pattern extraction algorithms.

Keywords: Text patterns, part-of-speech patterns, POS patterns.

1 Introduction

Text pattern matching has been used in Information Extraction [1], automated question [2] and e-mail [3] answering, and e-mail filtering [4]. Any text recognition task that involves regular expressions means matching surface text patterns. To the best of our knowledge, however, no one has studied what actually happens when text patterns do match pieces of text being recognized.

In this paper, we explore matching of manual text patterns to e-mail messages in order to assign standard answers, i.e., in order to put a message into a specific text category. The syntax of the text patterns resembles regular expressions. Content-wise the text patterns embody informal human perception of important wording for particular e-mail inquiries. In a text pattern, we can define a number of synonyms that designate a concept, we can define the order of the words and the distance between the words. Words are represented by word stems that match different inflections of a word. The text patterns are convenient for matching compound words, which are common in Germanic languages. A complete description of the text patterns is available in [3].

When a text pattern matches a piece of query text, it leaves a footprint. The difference between a text pattern and its footprint is the same as the difference between a regular expression and the bits of text it matches. We restrict a footprint to one sentence, which makes it easier to establish a representative number of words and

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 337–344, 2014.

the distance between the words in the footprints, as well as establish part-of-speech patterns in the footprints. In the following example, footprints are the underlined words: “I applied for housing allowance on 13 January 2009. When will my money come?” The two footprints are “applied housing allowance” and “when money come”.

We explore the footprints made by manual text patterns in e-mail messages during the process of automated e-mail answering. We explore by asking:

- How many words in query text do need to match a text pattern for successful matching outcome, i.e. correct message categorization?
- Are these words organized in n-grams or spread all around the sentence?
- How domain specific are the matching words?
- Do footprints form any part-of-speech (POS) patterns?

The language of the text being explored is Swedish, but we believe that similar conclusions would apply also to text in English.

The structure of this paper is straightforward. Section 2 describes the data that has created the footprints. Section 3 answers the above questions, and Sect. 4 summarizes the conclusions.

2 Experiment Data

The original collection was 9663 e-mail messages sent by citizens to the Swedish Social Security Agency. For our experiment, we selected 1909 messages, written in Swedish, that were correctly answered by an automated e-mail answering system, i.e. they were correctly assigned a standard answer, i.e., they were placed in a correct text category. 1882 messages belong to one text category while 27 messages belong to two categories (i.e. they have two standard answers), which makes 1936 message-category pairs. The categories and the number of selected messages in them are Cat1 (330), Cat2 (269), Cat3 (174), Cat4 (103), and Cat5 (1060). We ignored incorrect message categorization instances.

The size of the messages varied. The minimum, maximum, average, and median number of words per message were 4, 321, 45.5, and 35. The minimum, maximum, average, and median number of sentences were 1, 45, 5.2, and 4.

154 text patterns were used in order to categorize the messages. Because the text patterns were created manually, they are subjective, and so are their footprints. In order to give a somewhat objective picture of the footprints we describe them by the correctness of e-mail message categorization. The 154 text patterns categorized e-mail messages with precision about 90% and recall about 60% in the joint set of Cat1–Cat5. (The description of the measurements lies outside the scope of this paper.) We assume, without any proof, that footprints left by a different set of text patterns that achieve the same level of correctness of e-mail categorization, these different footprints would be similar.

Although we have 1936 message-category pairs, the system reproduced only 1918 of them because the system placed 18 messages, which belonged to two categories, into only one category.

3 Experiment Results

3.1 Amount of Matching Text

In total, 2273 sentences in the 1918 categorization instances matched one of the 154 text patterns, i.e. we have 2273 footprints. In 1577 categorization instances the message has one footprint (only one sentence in the message matched a text pattern), in 327 categorization instances there are 2 footprints, and in 14 categorization instances 3 footprints. The total number of words in the 2273 footprints is 12 108.

The overwhelming majority – 82% – of the messages in our experiment have only one matching sentence. We should bear in mind, however, that our messages lie within five specific text categories and have explicit information needs. Most information needs that were captured by the system happened to be stated in one sentence.

Table 1 shows the number of matching words per sentence (i.e., the size of a footprint) and per message, and the number of corresponding sentences and messages. The number of sentences in the table is independent from the number of messages. The majority of the messages – 1338 or 70% of the total – have 5 to 7 matching words. It is 5 to 7 times less than the median number of words per message, which is 35. For 19 messages, a correct answer was assigned on the grounds of only 3 matching words; these were very specific 3 words. In half of the messages, the footprints occupied less than 18% of the text.

Table 1. Number of matching words per sentence and per message

Matching words	Num. sentences	Num. messages	Matching words	Num. sentences	Num. messages
1	43	0	8	141	207
2	30	0	9	66	123
3	212	19	10	17	49
4	420	162	11	6	12
5	565	473	12	0	7
6	505	501	15	0	1
7	268	364			

3.2 Gaps between Matching Words

Knowing the number of matching words is not enough for having a complete idea of what the footprints look like. The distribution of the words across the sentence is another characteristic.

A gap between two matching words is a number of non-matching words between them. A gap of size 0 means that two matching words follow each other; a gap of size 1 means that there is a random word in between. Table 2 shows the gap sizes between matching words across all the footprints and the frequency of these gap sizes. In about 84% of the cases the gap is no larger than 1, which means that the matching words prefer staying in a cluster instead of evenly spreading across the sentence.

Table 2. Gap sizes across all the footprints

Gap size	Num.	%	Gap Size	Num.	Gap Size	Num.
0	6092	61.9	4	218	8	31
1	2155	21.9	5	106	9	17
2	728	7.4	6	76	More	56
3	309	3.1	7	47	Total	9835

An n-gram is a sequence of *n* words that follow each other; there is no gap in between. Table 3 shows the number of n-grams in all the footprints, as well as the number of distinct n-grams extracted from the footprints. The distinct n-grams are made of word lemmas and do not contain duplicates. The fourth column shows how many distinct n-grams could be decomposed into extracted distinct bi- and trigrams, while the fifth column permits also larger n-grams in the decomposition. Please observe that decomposition is not overlap. Our 9-grams could not be decomposed, but they certainly overlap with other n-grams.

Table 3. N-grams in footprints and extracted from footprints

Size	Num. in footprints	Num. distinct	Split into 2-3-grams	Split into 2-to-7-grams	
				Num.	%
1-gram	2981	428	n/a	n/a	n/a
2-gram	1374	544	n/a	n/a	n/a
3-gram	823	451	n/a	n/a	n/a
4-gram	468	282	62	62	22
5-gram	232	157	61	61	38.9
6-gram	102	74	7	15	20.3
7-gram	25	23	7	8	34.8
8-gram	8	8	1	5	62.5
9-gram	3	3	0	0	0
Total	**6016**	**1970**	**138**	**151**	

The number of distinct 2-to-9-grams is half the total number of 2-to-9-grams, which means the frequency of individual n-grams in the footprints is not high and they are not statistically representative. Because the text patterns, which made these n-grams, operate a rich vocabulary of synonyms, we have a good reason to believe that the n-grams would become more statistically representative if they were made of

concepts, which cover different synonyms, rather than word lemmas. This means that text pattern extraction from e-mail messages is likely to be more successful if these patterns are made of concepts, not individual words.

Our initial assumption was that larger n-grams would be easily decomposable into smaller n-grams, and we could cover the footprints with uni-, bi-, and trigrams. Our n-grams made of word lemmas show results different from what we hoped for.

3.3 Part-of-Speech Patterns

The POS pattern of a footprint is a sequence of POS attributes of the words in the footprint disregarding the gaps between the words. Two footprints may share the same POS pattern while containing different words. In order to discover the POS patterns, we did automatic POS-tagging of the entire email text.

Table 4. Most frequent POS patterns

No.	Swedish POS patterns	N	Examples translated to English	N
1	vb nn nn	51	sent application housing-allowance	17
2	pn ab vb nn	43	I/we not got [payment]	22
3	pn vb nn	41	I applied housing-allowance	19
4	nn	37	[domain specific concepts]	37
5	ha vb ps nn	35	when comes my [payment]	33
6	pn vb nn pp nn	27	I need form about parental-allowance	3
7	ha vb nn	26	when comes [payment]	10
8	vb ha pn vb nn	24	wonder when I get [payment]	18
9	vb nn ab vb nn	23	applied housing-allowance not got [reply]	22
10	pn ab vb nn pp nn	23	I not got [payment] in/for [period]	22
11	vb ha jj nn vb ab pp	22	wonder how many [days] left over for	22
12	ha vb pn ps nn	20	when get I my [payment]	18
13	vb pn vb nn	20	can/would you send form	14
14	pn vb vb nn	19	I want order form/brochure	14
15	vb ha ps nn vb	17	wonder when my [payment] comes	12
16	vb ab vb nn	17	have not got [domain-dependent-noun]	17
17	vb nn pp nn	16	sent papers about [allowance]	8
18	pn vb pn vb nn	15	I want you send form/brochure	13
19	vb ha nn vb	15	wonder when [payment] comes	12
20	vb ha jj nn vb vb pl pp	14	wonder how many [days] have taken out for	12

Our 2273 footprints host 942 POS patterns. Table 4 shows 20 most popular ones. The parts-of-speech in the table are verbs (vb), nouns (nn), pronouns (pn), question adverbs (ha), adverbs (ab), prepositions (pp), adjectives (jj), possessives (ps), and particles (pl).

The frequency of individual POS patterns drops quickly. 600 of the 942 POS patterns (63.7%) occur only once. Table 4 shows also the most representative example of each POS pattern. The examples are word lemmas translated to English. A lemma in square brackets stands for a number of close synonyms that represent the concept. For example, [days] means different wordings for child-care days paid by the state.

The POS patterns are dominated by individual expressions. For example, the pattern no. 11 occurs 22 times, and always with the same expression. The domination of individual expressions weakens the role of the POS patterns, detached from the text patterns, as representative features of the text categories.

Table 5. Most frequent lemmas

POS	Lemma	English	N	POS	Lemma	English	N
pn	jag	I	841	pp	för	for	198
vb	få	get	724	nn	*ersättning*	*compensation*	159
ha	när	when	467	vb	kunna	can	157
ab	inte	not	448	Jj	många	many	149
vb	undra	wonder	386	vb	vilja	want	147
vb	ha	have	371	nn	*dag*	*day*	142
ha	hur	how	334	vb	vara	be	137
vb	*skicka*	*send*	294	pp	på	on	126
nn	*blankett*	*form*	293	nn	*ansökan*	*application*	122
nn	*peng*	*money*	274	pn	det	this/that	111
ps	min	my	258	vb	ta	take	111
nn	*bostads-bidrag*	*housing allowance*	246	nn	*föräldrapenning*	*parental allowance*	157
vb	komma	come	244	pl	ut	out	94
nn	*utbetalning*	*payment*	226	nn	*pension*	*pension*	88
pn	ni	you	223	vb	*betala*	*pay*	82

Table 5 shows the most frequent words (their lemmas in Swedish and translation into English) in the footprints. Domain-specific words, in italic, designate what the e-mail inquiry was about. For example, “day” is essential for the concept of child-care days; “pay” is essential in expressions about payments.

Somewhat surprisingly, seven most frequent are common language words, not domain words. The footprints mirror manual text patterns crafted to embody the essence of an inquiry in an e-mail message. People do not communicate through sets of keywords; the words that help formulate intelligible sentences are as important as domain-specific keywords. Khosravi and Wilks [5] have observed that the share of nouns, verbs, adjectives, and adverbs in e-mail messages is about 40%, close to that in spoken language. More than half are words that support the communication.

3.4 POS Patterns across Text Categories

The distribution of the POS patterns in the footprints across Cat1–Cat5 is uneven (see Table 6). We are curious how representative the POS patterns are in the entire e-mail text across Cat1–Cat5. If certain POS patterns do stick out, we could use them in text categorization outside the scope of our text patterns matching. In order to judge representativeness of the POS patterns, we have to normalize their number per category with respect to the size of the category. Normalized POS pattern frequencies (not included here because of space limitations) show that most POS patterns are somewhat evenly distributed across Cat1–Cat5. Some irregularities are marked in Table 6: weak overrepresented in bold, strong overrepresented in bold underlined, underrepresented in italic underlined. POS pattern no. 7 covers a variety of expressions (see Table 4), and it is underrepresented in Cat1. The other six POS patterns with irregularities cover mostly one expression, their content lacks diversity. Interesting is Cat3: POS patterns no. 5, 12, 16, and 19 are not represented in the footprints but are overrepresented in the rest of the text.

Table 6. Number of most frequent POS patterns (referenced by their sequence number in Table 4) across text categories

No.	In the footprints, Cat1–Cat5					In the entire text, Cat1–Cat5				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	3	37	0	1	10	1643	1238	699	607	4585
2	12	15	0	0	16	406	396	166	127	1412
3	9	21	0	7	4	1197	982	529	381	3522
4	5	15	7	1	9	3270	2437	1420	1198	9377
5	0	0	0	0	35	*25*	65	**97**	22	326
6	15	5	0	1	6	603	494	294	205	1652
7	1	7	0	7	11	*97*	189	171	88	774
8	0	4	0	0	20	187	234	216	108	785
9	0	22	0	0	1	326	365	151	134	1230
10	0	0	0	0	23	165	145	84	57	553
11	0	0	22	0	0	20	*4*	**143**	*2*	45
12	0	2	0	0	18	13	25	**29**	11	127
13	15	4	0	0	1	1056	812	477	382	2944
14	17	0	0	1	1	857	672	357	260	2438
15	0	7	0	0	10	601	614	251	223	2220
16	0	0	0	0	17	47	39	**55**	**34**	237
17	2	1	0	0	13	1286	939	545	468	3349
18	15	0	0	0	0	460	391	240	158	1431
19	0	0	0	0	15	*109*	144	**288**	77	718
20	0	0	14	0	0	16	*3*	**73**	2	15

4 Conclusions

We have researched text pattern matching in order to categorize e-mail messages into specific text categories, where all messages in one category share the same standard answer. We have explored the traces of manual text patterns in the text of almost two thousand e-mail messages.

In 70% of all categorization instances, only 5–7 words in a message were used in order to decide the right text category for the message. Half of the messages were categorized using less than 18% of their text.

A gap between two words in an e-mail message that match a manual text pattern is the number of non-matching words between the matching words. About 84% of such gaps are no larger than 1; representative words tend to stick together.

About 75% of all matching words lie in n-grams of size 2 to 9. The number of distinct 2-to-9-grams, where duplicates are removed, is half of the total number of the 2-to-9-grams found in the message text, which means that individual n-grams are not statistically representative. We believe that n-grams made of concepts, which enclose a number of synonyms, rather than word lemmas would be more representative.

A POS pattern covers the words in one sentence that have correctly matched a manual text pattern. Such POS patterns are not representative features of the text categories. We do believe, however, that mixing text patterns and POS patterns in text pattern matching could increase the recall of categorization.

Unlike one would expect, most matching words are common language words. It is the combination of common words and a few domain keywords that is representative for an inquiry, not the single words themselves. If we were to extract inquiry-specific text patterns automatically, focusing on domain keywords or inverted document frequency would not help, except maybe for extracting seed terms the way Downey et al. [1] did.

References

1. Downey, D., Etzioni, O., Soderland, S., Weld, D.S.: Learning text patterns for web information extraction and assessment. In: Proc. AAAI 2004 Workshop on Adaptive Text Extraction and Mining, pp. 50–55 (2004)
2. Sneiders, E.: Automated FAQ Answering with Question-Specific Knowledge Representation for Web Self-Service. In: Proc. 2nd International Conference on Human System Interaction (HSI 2009), Catania, Italy, May 21-23, pp. 298–305. IEEE (2009)
3. Sneiders, E.: Automated Email Answering by Text Pattern Matching. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) IceTAL 2010. LNCS (LNAI), vol. 6233, pp. 381–392. Springer, Heidelberg (2010)
4. Wang, J.H., Chien, L.F.: Toward Automated E-mail Filtering – An Investigation of Commercial and Academic Approaches. In: TANET 2003 Conference, pp. 687–692 (2003)
5. Khosravi, H., Wilks, Y.: Routing email automatically by purpose not topic. Natural Language Engineering 5, 237–250 (1999)