Exploring the Traits of Manual E-Mail Categorization Text Patterns

Eriks Sneiders, Gunnar Eriksson, Alyaa Alfalahi
{eriks,gerik,alyalfa}@dsv.su.se

Stockholm University

Introduction

We explore matching of manual text patterns to e-mail messages in order to assign standard answers to routine questions sent to a customer service center. In essence, this is a text categorization task. Such text categorization can reach 90% precision at recall levels of around 60%.

We would like to know *how* the text-pattern matching occurs, *what happens* when a text pattern matches a piece of e-mail text. For example, a text pattern matches a number of relevant words in the query text: "I applied for housing allowance on 13 Jan 2009. When will my money come?" The words "applied", "housing", "allowance", "when", "money", "come" are the matching words whose characteristics we explore. We wonder:

- How many words in the query text need to match a text pattern for a successful matching outcome?
- Are these words organized in n-grams or spread all over the sentence?
- How domain specific are the matching words?
- Do the matching words form any POS patterns?

Answering these questions may help us improve text pattern extraction algorithms.

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 assigned a standard answer, i.e., they were placed in a correct text category. There were five standard answers and therefore five text categories Cat1 (330 messages), Cat2 (269 messages), Cat3 (174 messages), Cat4 (103 messages), and Cat5 (1060 messages).

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

The 154 text patterns, which we applied for text categorization, resemble regular expressions. An example of a text pattern:

```
[$me_we_etc :::; $need $acquire $apply $register :::::; $allowance
$subsidy $child_support]
[$me_we_etc :::; change changed changing move moved moving
::::::; bank $bank_account] ;

[can could $want shall should :::; [you $social_service ::::::::;
$send] [$me_we_etc ::::::::; $obtain] :::::; $fill_in_form]
```

The results section on the right shows some statistics around the matching words in the 1909 messages after a message had matched one of the 154 text patterns.

Conclusions

In about 70% of the text pattern-message matching cases, only 5-7 words in the query message are needed in order to decide which standard answer, if any, should be assigned to the message. In half of the cases, less than 18% of the message text is used.

About 84% of the gaps between the matching words, when a text pattern matches a query message, 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. Distinct 2-to-9-grams are made of word lemmas, duplicates are removed from the n-gram set. Diversity of these n-grams is large; individual n-grams are not statistically representative. Also, the n-grams are not easily decomposable into smaller n-grams from the same n-gram set. We believe that "n-grams" made of matching concepts, where each concept covers a number of matching synonyms, would be more statistically representative. Also, an "n-gram" made of matching concepts would, hopefully, be easier to decompose into smaller n-grams. Therefore automatic text-pattern extraction is more likely to yield generic patterns if it extracts patterns of concepts rather than patterns of word lemmas.

POS patterns of matching words are not representative features of 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. While domain-specific words define the topic of an e-mail inquiry, common language words define the question, complaint, or request itself.

Results

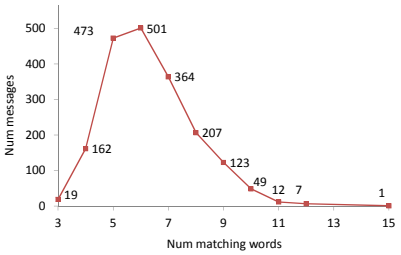

Figure 1 shows the number of words in an e-mail message that a text pattern matches in order to assign a standard answer. The majority of the messages – 1338 or 70% of the total – have 5 to 7 matching words. It is 5 to 7 times less than the median number of words per message, which is 35.

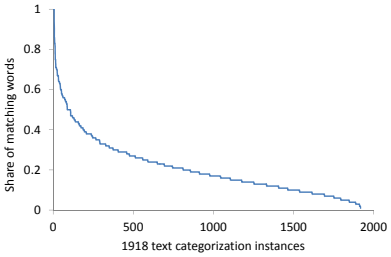

Figure 2. Share of the matching words among all the words in the text of the query message across 1918 text categorization instances. In half of the messages, the matching words occupied less than 18% of the message text.

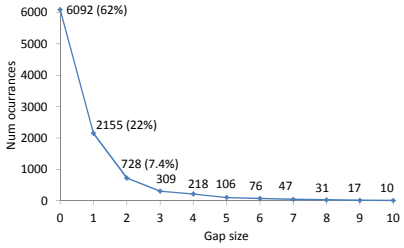

Figure 3. A gap between two matching words is a number of non-matching words between them. The graph shows eleven most frequent gap sizes and the frequency of each gap size. In about 84% of the cases the gap is no larger than 1, which means that the matching words prefer staying in a cluster instead of evenly spreading across the sentence.

Size	Num. in footprints	Num. distinct	Split into 2-3-grams	Split into 2-to-7-grams Num.	Split into 2-to-7-grams %
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	

Figure 4 shows the number of n-grams made of the matching words, as well as the number of distinct n-grams, made of word lemmas, after duplicates have been removed. 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.

The number of distinct 2-to-9-grams is half of the total number of 2-to-9-grams, which means that individual n-grams are not statistically representative. The share of decomposable distinct n-grams is 22-39% in the majority of cases.

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

Figure 5. The POS pattern of the matching words is a sequence of POS attributes of these words in the sentence disregarding the gaps between these words. The table shows 20 most frequent POS patterns. The frequency of individual POS patterns drops quickly. 600 of the 942 POS pat-terns (63.7%) occur only once.

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

No.	Normalized in Cat1–Cat5					Original in Cat1–Cat5				
	1	2	3	4	5	1	2	3	4	5
1	115	106	93	136	100	1643	1238	699	607	4585
2	92	111	72	93	100	406	396	166	127	1412
3	109	110	92	111	100	1197	982	529	381	3522
4	112	102	92	131	100	3270	2437	1420	1198	9377
5	25	79	**181**	69	100	25	65	97	22	326
6	117	118	108	128	100	603	494	294	205	1652
7	40	96	135	117	100	97	189	171	88	774
8	77	117	168	142	100	187	234	216	108	785
9	85	117	75	112	100	326	365	151	134	1230
10	96	103	93	106	100	165	145	84	57	553
11	143	35	**1936**	46	100	20	4	143	2	45
12	33	78	139	89	100	13	25	29	11	127
13	115	109	99	134	100	1056	812	477	382	2944
14	113	109	89	110	100	857	672	357	260	2438
15	87	109	69	103	100	601	614	251	223	2220
16	64	65	**141**	**148**	100	47	39	55	34	237
17	123	110	99	144	100	1286	939	545	468	3349
18	103	108	102	114	100	460	391	240	158	1431
19	49	79	**244**	110	100	109	144	288	77	718
20	343	79	**2965**	137	100	16	3	73	2	15

Figure 6 shows the frequency of our top 20 POS patterns in Cat1 through Cat5. On the right side we see the actual numbers of occurrences of each POS pattern in each text category. On the left side we have normalized POS pattern frequencies; normalized with respect to the size of each text category and scaled to always 100 occurrences in Cat5.

The normalized frequencies show that most POS patterns are somewhat evenly distributed across Cat1–Cat5. Infrequent irregularities do occur.

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

Figure 7. The most frequent matching words, their lemmas in Swedish and the translation into English. Domain-specific words, in italic, designate the context of the inquiry, its subject area. Somewhat surprisingly, seven most frequent are common language words, not domain words. People do not communicate through sets of keywords; the words that help formulate an intelligible inquiry are essential.