Business Listings in Automatic Directory Assistance

Odette Scharenborg, Janienke Sturm, Lou Boves

A^{2}RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{O.Scharenborg,Janienke.Sturm,L.Boves}@let.kun.nl

Abstract

So far most attempts to automate Directory Assistance services focused on private listings, because it is not known precisely how callers will refer to a business listings. The research described in this paper, carried out in the SMADA project, tries to fill this gap. The aim of the research is to model the expressions people use when referring to a business listing by means of rules, in order to automatically create a vocabulary, which can be part of an automated DA service.

In this paper a rule-base procedure is proposed, which derives rules from the expressions people use. These rules are then used to automatically create expressions from directory listings. Two categories of businesses, viz. hospitals and the hotel and catering industry, are used to explain this procedure. Results for these two categories are used to discuss the problem of the over- and undergeneration of expressions.

1. Introduction

Directory Assistance (DA) is one of the telecom services with the largest call volumes. Although an increasing proportion of DA requests is now handled by self-help Internet services, the need for assistance by human agents will remain substantial, if only because of the increasing call volume from mobile networks, where callers do not have ready Internet access. The cost of human agents in a service with small added value like DA is becoming prohibitive. Therefore, virtually all major telecom operators are searching for ways in which DA services can be automated.

Intuitively it seems attractive to focus attempts to automate DA services on the most frequently called numbers. It has long been known that there is a strong correlation between the number of times a telephone number is called and the frequency with which that listing is the subject of a DA call [1,2]. It takes only a moment thinking to realise that these are the numbers of large companies, services and administrations. Unfortunately, it also appears that it is difficult to predict how callers will refer to these entities when they call the DA service. This lack of knowledge of caller behaviour seems to have been the major reason why most previous research in automating DA services has focused on residential listings. The SMADA project (Speech Driven Multi-modal Automatic Directory Assistance) is squarely aimed at business listings. The research reported in this paper aims to find ways to predict the expressions people use to refer to business listings.

The approach presented in this paper attempts to model expressions of frequently requested business listings by means of rules. For practical purposes we intend to deploy an isolated word recogniser. This forces us to consider each expression as a (compound) word. Thus, we need to develop a vocabulary that describes the way in which callers express the requests for frequently requested listings. The methodology with which the problem is tackled is based on a combination of knowledge and data.

This paper is organised as follows. Section 2 describes the material that was used. Section 3 explains the rule-based approach. Section 4 gives detailed information about the obtained lexicons and discusses the problems of over- and undergeneration. In section 5 a general discussion is presented. Finally, in section 6 our conclusions are drawn.

2. Material

For our research we used data from two sources, viz. the official directory database and the Dutch Directory Assistance Corpus (DDAC2000) [3]. This corpus consists of speech material obtained from calls to the Dutch 118 Directory Assistance service. The customers completed a dialogue in which they were prompted for the name of the city, the name of the company or person, and the street name. All answers were recorded and orthographically transcribed by humans. This research focuses on the expressions people used when answering the second question. The DDAC2000 corpus consists of 45,613 calls in which the question regarding the name of the business or person was answered. About 80% of the calls pertain to business listings.

Business listings of different categories of companies differ substantially from each other. Therefore, separate category-based vocabularies and recognition grammars have to be developed. The research described in this paper concentrates on the categories 'hospitals' (HOSP) and 'hotel and catering industry' (HCI). These two categories are chosen because (at first sight) they seem easy to define and they make up a substantial part of the calls to business listings.

For both categories a subcorpus was extracted from the DDAC corpus. For the HOSP category all utterances were selected that contain the keywords ' ziekenhuis' (E: ' hospital') or 'g asthuis' (a synonym for ' ziekenhuis'). The HOSP-corpus thus selected contains 428 utterances. The same keyword strategy, using the words *cafe, restaurant, hotel, pizzeria, cafetaria* and *snackbar* (synonym for *cafetaria*) was used to extract the HCI-corpus. This subcorpus contains 927 utterances. Together, the HOSP and HCI subcorpora make up about 3% of the DDAC2000 material.

To find a telephone number the corresponding listing must be retrieved from the directory database. Therefore, the listings in that database are the basis of the expressions we want to model. Several versions of this database exist, optimised for computer-assisted human search. These representations are not necessarily optimal for a non-intelligent ASR interface. Many businesses and organisations have multiple entries for a single telephone number.

3. Procedure

To generate a list of expressions for each category we first analysed the utterances in the two subcorpora to obtain a simple grammar that covers as many of the recorded expressions as possible. The aim is then to derive transduction rules, which convert the directory entries into the expressions used by the customers (and covered by the grammar). This procedure is explained in some detail for the categories HOSP and HCI.

3.1. Descriptive rules

All utterances in the subcorpora were described by means of rules. Rules that only describe one utterance were discarded. After the introduction of variables and generalisation of the rules the following four descriptive rules were obtained for HOSP:

1. (<type of>) *ziekenhuis* (<name>|<city>)
2. <name> *ziekenhuis* (<city>)
3. <name> *gasthuis*
4. (<name>) <prefix>*ziekenhuis* (<city>)

For HCI, the following five rules were obtained:

1. <est 1> <name>
2. <name> <est 2>
3. (<type of>) *restaurant*|*hotel* (<name>)
4. <prefix>*restaurant*|*cafe* <name>
5. (<name>) <prefix>cafe

Words between angled brackets are variables, parentheses indicate optional[1] elements and the slash '|' indicates an 'or'. Words in italics are terminals. The following variables need some more explanation:

- <type>: refers to the type of hospital, viz. *academisch* (E: academic) and *psychiatrisch* (E: psychiatric) or to the type of hotel or restaurant, e.g. Chinese, Italian, etc..
- <prefix> : refers to the type or name of a hospital or establishment, that is (unlike the <type> variable) is attached to the terminal to for a compound, e.g. *kinderziekenhuis* (E: ' children's hospital') or *nachtcafe* (E: ' night pub').
- <est 1> and <est 2>: refer to different sets of words that are used to name the different types of establishments, viz. cafe, restaurant, hotel, snackbar cafetaria, pizzeria, cafe restaurant, hotel restaurant for <est 1> and cafe, restaurant, hotel, snackbar, pizzeria for <est 2>. All words from <est 1> occur in front of the name of the establishment, but only a subgroup of these words (<est 2>) occurs behind the name.

The HOSP rules cover 99.5% of the HOSP-corpus, while the HCI rules describe 98.2% of the HCI-corpus.

3.2. Transduction rules

From the descriptive rules we generated transduction rules, that convert the directory listings into expressions used by customers. First, all listings for hospitals were retrieved from the country-wide directory database, yielding a total of 657 records for HOSP. Listings for the HCI category were retrieved from the directory database pertaining to Rotterdam (the second largest city in the Netherlands) and its immediate environment. This yielded 1,124 distinct directory listings for HCI.

The procedure for the generation of expressions callers might use works as explained by means of an example, viz. the directory listing '*ijsselland ziekenhuis*'. This listing is first parsed by the HOSP grammar, yielding '*ziekenhuis*' as terminal and '*ijsselland*' as a value of the variable <name>. '*ziekenhuis*' matches the conditions of rule 1 as well as those of rule 2. Next, all matching rules are used to generate expressions. This returns '*ziekenhuis ijsselland*' as output from rule 1 and '*ijsselland ziekenhuis*' as output of rule 2.

4. Results

Applying the procedure described in the previous section, lists of expressions were generated for both categories. For HOSP 967 distinct expressions are obtained and 2,094 for HCI (cf. Table 1).

	HOSP	HCI
# Hospitals or establishments	301	937
# Orthographically distinct directory listings	657	1,124
# Generated expressions	967	2,094
% Maximum achievable coverage in DDAC2000	99.5 %	98.2 %

Table 1: Overview of the number of listings and the number of generated expressions.

Since more than one rule can apply to a directory listing; several expressions per directory listing can be generated. As can be seen in Table 1, hospitals have an average of 2.2 distinct directory listings, while the average number for a cafe, hotel or restaurant is only 1.2 (#Orthographically distinct directory listings / #Hospitals or establishments). One of the reasons for this difference could be that hospitals should be easy to find in case of emergencies. Therefore, it is recommendable to use several distinct directory listings. On the other hand, not immediately finding the telephone number of a cafe, hotel or restaurant is surmountable; therefore, a single directory listings may be sufficient.

Since directory listings belonging to the same hospital most of the time only vary in the order of the words (for example '*ziekenhuis ijsselland*' and '*ijsselland ziekenhuis*'), the HOSP rules in section 3.1 generate identical expressions for both entries in the directory database. This is much less the case for HCI. Consequently, the average number of generated expressions per distinct directory listing (#Generated expressions / #Orthographically different listings) is only 1.5 for HOSP and 1.9 for HCI. However, from looking at the number of distinct hospitals and HCI establishments relative to the number of generated listings (#Generated expressions / #Hospitals or establishments), it appears that the average number of expressions generated for HOSP is higher than for HCI, 3.2 and 2.2 respectively. This is easily explained by looking at the descriptive rules; there are more words in a HOSP directory listing and therefore more variables in the rules. More variables increases the number of generated expressions.

[1] In rules with several optional elements at least one of the optional elements should be present

Rule	#Generated expressions		#Found expressions		Overgenera-tion		Coverage of corpus	
	#	%	#	%	#	%	#	%
HOSP								
(<type of>) *ziekenhuis* (<name>) | (<city>)	382	39.5	35	9.2	347	90.8	85	19.8
<name> *ziekenhuis* (<city>)	393	40.6	51	13.0	342	87.0	185	43.1
<name> *gasthuis*	77	8.0	5	6.5	72	93.5	9	2.1
(<name>) <prefix>*ziekenhuis* (<city>)	115	11.9	15	13.0	100	87.0	52	12.1
Total	967	100	106	11.0	861	89.0	331	77.2
HCI								
<est 1> <name>	931	44.5	13	1.4	918	98.6	26	31.7
<name> <est 2>	771	36.8	2	0.3	769	99.7	2	2.4
(<type>) *restaurant* / *hotel* (<name>)	212	10.1	3	1.4	209	98.6	4	4.9
<prefix>*restaurant* / *cafe* <name>	92	4.4	0	0	92	100.0	0	0
(<name>) <prefix>*cafe*	88	4.2	2	2.3	86	97.7	2	2.4
Total	2094	100	20	0.1	2074	99.9	34	41.5

Table 2: The absolute number of generated expressions per rule, their coverage and overgeneration and the accompanying percentages per rule and the number of covered expressions in the corpus.

Table 2 gives more detailed information on the composition of the two lexicons that were generated from the entries in the directory database by the rules in 3.1.

The second (broad) column of Table 2 shows for each rule the (absolute and relative) number of expressions that were generated by applying this rule to the directory database. The third column (labelled '#Found expressions') shows how many of these generated expressions (types) were actually found in the DDAC2000. The fourth column (labelled 'Overgeneration') shows the overgeneration, i.e. the number and proportion of expressions that were generated but never occurred in the subcorpus. Finally, the last column (labelled 'Coverage of corpus') gives the number and proportion of utterances in the DDAC2000 subcorpus that matches one of the generated expressions (tokens).

In the case of HCI, the numbers mentioned in Table 2 are not based on the same corpus we used for the formulation of the rules. Since we only used the directory listings for HCI in the city of Rotterdam we could only generate expressions for the HCI of Rotterdam. Thus, we used a subset of the HCI-corpus to estimate the coverage and overgeneration, viz. only the expressions of establishments located in Rotterdam are considered.

As can be seen from Table 2, 89.0% of the expressions that were generated for HOSP, were never used. The remaining 106 expressions cover 77.2% of the HOSP-corpus. This is less than the maximum achievable coverage of 99.5%. In the case of HCI 99.9% of the expressions were never used. The remaining 20 expressions cover only 41.5% of the corpus. The reason for this overgeneration and undergeneration will be further discussed below. As is shown in Table 2, only 11% of the generated expressions for HOSP can be found in its subcorpus and only 0.1% of the expressions for HCI.

It has to be kept in mind that for HOSP 967 expressions are generated, while there are only 429 corpus utterances. The DDAC2000 corpus does not contain requests for each of the 301 hospitals in the country. Thus, our corpus is evidently too small to allow for simple conclusions. In case of HCI the maximum coverage of expressions is only 3.9%, since there are 2,094 expressions generated, while the DDAC2000 subcorpus for the city of Rotterdam only consists of 82 utterances. Yet, a detailed analysis of the generated expressions that were not found in the DDAC2000 subcorpora has revealed a number of problems that cause 'hard' overgeneration.

4.1. Overgeneration

The overgeneration is determined by matching all generated expressions with the transcriptions in the category specific subcorpora. If an expression does not occur in the subcorpus, it is said that it is overgenerated.

One cause of hard overgeneration for HOSP is due to the keyword approach that was used to select the directory listings referring to a hospital. With the keyword method directory listings can be selected that do not really belong to HOSP. For instance, the listing '*Stichting kind en ziekenhuis Hengelo*' contains the keyword *ziekenhuis*, but actually refers to a welfare foundation instead of a hospital. For these listings expressions are generated as well, but the rules of section 3.1 do not apply here. A solution for this problem is to use a version of the directory database that contains information about the type and category of the listings, which would make the keyword-based search superfluous. Also, the use of optional variables in the rules inflates the number of generated expressions. The number of generated expressions doubles with each optional variable, since there is always one variant with and one without the variable.

For the HCI category the overgeneration is worse than for HOSP. Of course, the problems mentioned above also hold for this category, but there are two additional issues. First, the variables `<est 1>` and `<est 2>`, which only occur in the rules for HCI, can take a large number of values. One has not only 'singular' values like *cafe*, or *restaurant* appearing in the directory listing, but also compound values like *cafe restaurant, hotel cafe restaurant,* etc. exist. All of these 'singular' values and combinations are used to generate expressions. This process evidently causes overgeneration

The second problem concerns the number of different values for the variable `<type of>`. Whereas for HOSP only two different types were distinguished (cf. section 3.1), for HCI a lot of types exist, like ' Chinese' , 'fast-food', 'Indian', etc. and combinations of these. Directory listings often contain more than one type. Generating expressions using all types strongly increases the number of generated expressions.

4.2. Undergeneration

The last column of Table 2 shows that the generated expressions cover 77% of the HOSP corpus and 42% of the HCI corpus. This is much lower than the maximum achievable values of 99.5% and 98.2% (cf. section 3.1). For HOSP undergeneration is due to the fact that the descriptive rules in section 3.1 appear to be not fully adequate as transduction rules. For example, callers who need the number of the hospital in a small town, where there is only one, often refer to this entity by '*het ziekenhuis*' (E.: the hospital). Since the database always has the hospital(s) by the full name, the short expression '*ziekenhuis*' will not be generated. Of course, this short expression is ambiguous for requests pertaining to larger towns, which have multiple hospitals. Thus, it appears that we need different rule sets, depending on the city. This problem is not unique for hospitals; it also affects airports, department stores, etc., which will be unique in small towns, but not necessarily also in bigger ones.

For the category HOSP this problem is complicated by the fact that callers are known to refer to hospitals by their name only. This is especially common for the large academic hospitals, which go under acronyms like AMC (*Academisch Medisch Centrum*) in Amsterdam. If the city is Amsterdam, the response '*het AMC*' to the prompt for the business name identifies a unique listing. However, not all hospitals are indicated in this way, so we again face the trade-off between over- and undergeneration. Actually, it is quite possible that we have missed requests for hospitals in DDAC2000 because the caller used the name without a reference to the fact that the entity represented a hospital.

In the case of HCI undergeneration is partly due to the fact that establishments in the directory listings are indicated in unusual manners (e.g. *eten en drinken* (E: food and drinks), *'mangerie'* (E: eatery), etc. If these directory listings also contain one of the keywords (for example, '*café eten en drinken de kok*') expressions with 'café' are generated. But since '*eten en drinken*' is not known by our grammar as a synonym of 'café' it is included in the `<name>` part of the expression. Thus, the expression '*café de kok*' (used by the caller) is not delivered; instead, the unlikely form '*café eten en drinken de kok*' appears. Therefore, these fancy descriptions increase the undergeneration

Another instantiation of the same problem with the HCI category is the fact that people often use a type of establishment in their utterance that is different from the one in the directory. For instance, callers often ask for *restaurant New York*, while the directory database only has *hotel New York*. Our transduction rules do not generate the expression *restaurant New York*, again to prevent wild overgeneration.

5. General discussion

One problem with business listings that remains to be solved pertains to the fact that callers use expressions that can in no way be derived from the text in the listings database. 'Nicknames', viz. *bijlmerbajes* instead of *penitentiaire inrichtingen over-amstel* (the big prison in Amsterdam), are especially cumbersome. We intend to approach this problem by interviewing senior call centre agents, who most probably have lists of this type of 'oddities'.

Especially in the case of HCI, the figures presented in this paper are difficult to interpret; there is too few data available to make hard statements of whether this approach is trustful. The lexicon rapidly grows when all possible expressions are added. The problem is bigger with the 'hotel and catering industry' category than with the 'hospital' category. One way to prevent uncontrollable overgeneration would be to do the recognition in two steps: first scan the response for a small number of keywords, and then process the same utterance with one or more specialised recognisers, each tuned to what one would expect for the keywords.

In recording the DDAC2000 corpus the numbers that were released by the agents could not be included. Consequently, the transcribers had to rely on their acoustic recognition of the names, without the support of an intelligent database search algorithm. This caused numerous transcription errors, many of which were extremely difficult to detect. In future research DA calls should only be recorded if the released number can be included, so that reverse directory information can be used to support the transcription.

6. Conclusion

In this paper a rule-base procedure is proposed, that derives rules from the expressions people use, when asking for a directory listing of a company. These rules were then used to automatically create expressions from directory listings. It turned that the generated expressions for the categories 'hospitals' and 'hotel and catering industry' were not able to completely cover all expressions found in the DDAC2000 material. Expressions were both under- and overgenerated. The causes and some solutions to these problems were presented.

The results suggest that a rule-based approach is a promising method to convert directory listing into expressions used by people. However, additional knowledge sources must be developed to tackle problems with unusual listings. It appears that different categories of businesses require different knowledge sources and different rules.

7. Acknowledgements

This research is supported by the EC under the IST-HLT Programme.

8. References

[1] Billi, R., Canavesio, F., Rullent, C. (1998). Automation of Telecom Italia Directory Service: Field Trial Results. Proc. 4th IEEE Workshop *Interactive Voice Technology for Telecommunications Applications*, pp. 11-16.

[2] Gupta, V., Robillard, S. and Pelletier, C. (1998). Automation of Locality Recognition in ADAS Plus. Proc. IEEE 4th Workshop *Interactive Voice Technology for Telecommunications Applications*, pp. 1-4.

[3] J. Sturm, H. Kamperman, L.Boves, E. den Os, Impact of speaking style and speaking task on acoustic models, *Proceedings ICSLP 2000*, p. 361-364, 2000.