# Seed Words Based Data Selection for Language Model Adaptation

**Roberto Gretter**, **Marco Matassoni**, **Daniele Falavigna**
Fondazione Bruno Kessler, Trento, Italy
(gretter,matasso,falavi)@fbk.eu

**Abstract**

We address the problem of language model customization in applications where the ASR component needs to manage domain-specific terminology; although current state-of-the-art speech recognition technology provides excellent results for generic domains, the adaptation to specialized dictionaries or glossaries is still an open issue. In this work we present an approach for automatically selecting sentences, from a text corpus, that match, both semantically and morphologically, a glossary of terms (words or composite words) furnished by the user. The final goal is to rapidly adapt the language model of an hybrid ASR system with a limited amount of in-domain text data in order to successfully cope with the linguistic domain at hand; the vocabulary of the baseline model is expanded and tailored, reducing the resulting OOV rate. Data selection strategies based on shallow morphological seeds and semantic similarity via word2vec are introduced and discussed; the experimental setting consists in a simultaneous interpreting scenario, where ASRs in three languages are designed to recognize the domain-specific terms (i.e. dentistry). Results using different metrics (OOV rate, WER, precision and recall) show the effectiveness of the proposed techniques.

## 1 Introduction

In this paper we describe an approach to adapt the Language Models (LMs) used in a system designed to give help to simultaneous interpreters. Simultaneous interpreting is a very difficult task that requires a high cognitive effort especially to correctly translate parts of the source language that convey important pieces of information for the final users. These are: numerals, named entities and technical terms specific of each interpretation session. As an example, a study reported in Desmet et al. (2018) claims that the error rate made by professional interpreters on the translation of numbers is, on average, equal to 40%.

This demands for a technology, based on automatic speech recognition (ASR), capable of detecting, in real time and with high accuracy, the important information (words or composite terms) of a speech to interpret and to provide it to a professional interpreter by means of a suitable interface. Therefore, our goal is not to minimise the word error rate (WER) of an audio recording, as usual in ASR applications, instead we aim to maximise the performance of the developed system, in terms of precision, recall and F-measure, over a set of "important" terms to recognise, as will be explained in section 4. To do this we experimented on a set of data properly labelled by human experts.

It is worth to point out that this task is different from the usually known "keywords spotting" task, since we cannot assume to know in advance the terms to spot inside the audio stream but we can only start from some "seed" terms belonging to a glossary which is part of the experience of each human interpreter. This demands for further processing modules that: *a)* extend, in some way, the given glossary including also ”semantically” similar terms, as will be explained in section 2.2, in order to adapt both the dictionary and the language model (LM) employed in the ASR system, and/or *b)* detect along an automatically generated transcription the pieces of information (i.e. numerals, named entities, etc) useful to the interpreter. Actually, the ASR system described below is part of a bigger system that integrates natural language processing (NLP) modules, dedicated to both named entity and numeral extraction, and a user interface specifically designed according to the requirements of professional interpreters. This system, named SmarTerp[1], aims to support the simultaneous interpreters in various phases of their activities: the preparation of glossaries, automatic extraction and display of the “important” terms of an interpreting session, post-validation of new entries (Rodríguez et al., 2021).

**Related works.** As previously mentioned spotting known words from audio recordings is a largely investigated task since the beginning of speech recognition technology (e.g. see works reported in Bridle (1973); Rose and Paul (1990); Weintraub (1995). Basically all these approaches used scores derived from acoustic log-likelihoods of recognised words to take a decision of keyword acceptance or rejection.

More recently, with incoming of neural networks, technologies have begun to take hold based on deep neural networks (Chen et al., 2014), convolutional neural networks (Sainath and Parada, 2015) and recurrent neural networks (Fernandez et al., 2007) to approach keyword spotting tasks. The last frontier is the usage of end-to-end neural architectures capable of modelling sequences of acoustic observations, such as the one described in Yan et al. (2020) or the sequence transformer network described in Berg et al. (2021).

The approach we use for enlarging the dictionary of the ASR system and to adapt the corresponding language model to the application domain is to select and use from a given, possibly very large and general text corpus, the sentences that exhibit a certain “similarity” with the terms included in the glossaries furnished by the interpreters. Similarly to the keyword spotting task, “term based similarity” represents a well investigated topic in the scientific community since many years. A survey of approaches can be found in the work reported in Vijaymeena and Kavitha (2016). Also for this task the advent of neural network based models has allowed significant improvements both in the word representation, e.g. with the approaches described in Mikolov et al. (2013), and in text similarity measures, e.g. as reported in Le and Mikolov (2014); Amin et al. (2019).

Worth to notice is that in the ASR system used for this work we do not search for new texts to adapt the LM, instead, as explained in section 2, we select the adaptation texts from the same corpus used to train the baseline LM. Note also that our final goal is not that to extract the named entities from the ASR transcripts - this task is accomplished by the NLP modules mentioned above - instead it consists in providing to the ASR system a LM more suitable to help the human interpreter of a given event. Also for ASR system adaptation there is an enormous scientific literature, both related to language models and to acoustic models adaptation; here we only refer some recent papers: Song et al. (2019) for LM adaptation and Bell et al. (2021) for a review of acoustic model adaptation approaches, especially related to neural models.

## 2 Automatic selection of texts

Usually a Language Model (LM) is trained over huge amounts of text data in a given language, e.g. Italian. During the training phase, a fixed lexicon is selected - typically the N most frequent words in the text - and millions or billions of n-grams are stored to give some probability to any possible word sequence. This process allows to build a somehow generic LM, capable to represent the language observed in the text.

[1]The SmarTerp Project is funded by EIT DIGITAL under contract n. 21184

However, interpreters often need to specialise their knowledge on a very specific topic, e.g. dentistry. In this case, they also have to quickly become experts in that particular field. We could say that they need to adapt their general knowledge to that field: this means that, before the event, they have to collect material about that topic, study it, prepare and memorise a glossary of very specific technical terms together with their translations.

The same process holds for an ASR system: it can perform in a satisfactory way in a general situation, but it may fail when encountering technical terms in a specific field. So, it has to be adapted, both in terms of lexicon (it may be necessary to add new terms to the known lexicon) and in terms of word statistics for the new terms.

In the SmarTerp project we are going to explore different adaptation procedures and describe in this paper our preliminary work in this direction. At present, we hypothesise that an interpreter could provide some text and the ASR system will be able to adapt to the corresponding topic in a short time (some hours on a medium computer). This short text could range from a few words to a quite large set of documents that identify that particular topic, depending on the expertise and the attitude of the interpreter. Here are some possibilities:

- just a few technical words;
- a glossary of terms, maybe found with a quick internet search;
- a glossary of technical terms with translations, maybe built over the years by an expert interpreter;
- a set of technical documents, in the desired language.

In a very near future, in SmarTerp a pool of interpreters will be engaged in simulations where they have to provide data that, in a complete automatic way (i.e. without the intervention of some language engineer), will adapt the ASR system for a particular topic. In this work we are testing some tools and procedures in order to provide them some possible solutions, assuming that at least some small text (i.e. a glossary, or even a few words) will be available. From this small text we will derive some *seed words* that will be used, in turn, both to update the dictionary of the ASR system and to select LM adaptation texts from the available training corpora (see Table 2). In detail, we implemented the following procedures (although some of them were not used in the experiments described in this paper):

- selection of **seed words**, i.e. technical words that characterise the topic to be addressed; they are simply the words, in the short text provided by the interpreter, that are not in the initial lexicon, composed of the most frequent N words of that language (128 Kwords, in this paper).
- optional enlargement of the set of **seed words**, either by exploiting shallow morphological information or using neural network approaches like word2vec (Mikolov et al., 2013).
- selection of **adaptation text**, i.e. text sentences in the text corpus that contain at least one of the seed words. Note that we hypothesise not to have new texts belonging to the topic to be addressed, that could be directly used for LM adaptation.
- compilation of an **adapted lexicon** and of an **adapted LM**, obtained exploiting the adaptation text.

### 2.1 Shallow morphological seed words enlargement

Each initial seed word is replaced by a regular pattern which removes the ending part, to find similar words in the complete dictionary of the corpus. Possible parameters are: $N_M$, maximum number of similar words retained for each seed; $L_M$, minimal length of a seed pattern to be considered valid (too short patterns are useless or even dangerous).

| Language | CV (h:m) | EuroNews (h:m) | Total Speakers | Running words |
|---|---|---|---|---|
| English | 781:47 | 68:56 | 35k | 5,742k |
| Italian | 148:40 | 74:22 | 9k | 1,727k |
| Spanish | 322:00 | 73:40 | 16k | 2,857k |

Table 1: Audio corpora for AM training

### 2.2 Semantic similarity based approach

Each initial seed word is fed to a pretrained neural skipgram model (word2vect, see http://vectors.nlpl.eu/repository), which returns an embedded representation of words. Then, the $N$ more similar words are computed using the cosine distance between couples of words embeddings. The process can be iterated by feeding word2vec with every new similar word obtained. Possible parameters are: $N_W$, number of retained words from each term; $I_W$, number of iterations: typically 1, or 2 in case of a very short list of initial seeds.

### 2.3 Selection of adaptation text

Given a final set of seed words, the huge text corpus is filtered and every document containing at least one seed word, not contained in the (128K) initial lexicon, is retained. One parameter of the filter - not used in this work - is the number of words forming the context around every seed word in a document. This may be useful to avoid to include in the adaptation corpus useless pieces of texts, due to the fact that every line in the training corpora (newspaper or Wikipedia, title or article) is considered a document, containing from few words to tens (even hundreds in few cases) of Kwords. Note that the selection of the adaptation text is largely responsible of the lexicon enlargement (up to 250 Kwords, see Table 6), since the number of seed words resulted to be, in our preliminary experiments, always below 4 Kwords.

## 3 ASR systems

The ASR system is based on the popular Kaldi toolkit (Povey et al., 2011), that provides optimised modules for hybrid architectures; the modules support arbitrary phonetic-context units, common feature transformation, Gaussian mixture and neural acoustic models, n-gram language models and on-line decoding.

### 3.1 Acoustic models

The acoustic models are trained on data coming from CommonVoice (Ardila et al., 2020) and Euronews transcriptions (Gretter, 2014), using a standard *chain* recipe based on lattice-free maximum mutual information (LF-MMI) optimisation criterion (Povey et al., 2016). In order to be more robust against possible variations in the speaking rate of the speakers, the usual *data augmentation* technique for the models has been expanded, generating time-stretched versions of the original training set (with factors 0.8 and 1.2, besides the standard factors 0.9 and 1.1).

Table 1 summarises the characteristics of the audio data used for the models in the three working languages considered in the project.

### 3.2 Language models and Lexica

Text corpora that can be used to train LMs for the various languages are described in Table 2 and derive both from Internet news, collected from about 2000 to 2020, and from a Wikipedia dump; their corresponding total lexica amount to several millions of words (from 4 to 10) for every language. It has to be clarified that, being the original texts definitely not clean, most of

| Language | Lexicon size | Total running words | Internet News | Wikipedia 2018 |
|---|---|---|---|---|
| English | 9.512.829 | 3790.55 Mw | 1409.91 Mw | 2380.64 Mw |
| Italian | 4.943.488 | 3083.54 Mw | 2458.08 Mw | 625.46 Mw |
| Spanish | 4.182.225 | 2246.07 Mw | 1544.51 Mw | 701.56 Mw |

Table 2: Text corpora for training the LMs for ASR in the three SmarTerp languages. Mw means millions of running words.

the low frequency words are in fact non-words (typos, etc.). For practical reasons, the size of the lexicon used in the ASR usually ranges from 100 to 500 Kwords.

The baseline language models are trained using the huge corpora described in Table 2; the adaptation set is selected from the same huge corpora. After the selection stage, the resulting trigrams are computed and a mixed LM is built and then pruned to reach a manageable size. The adapted LM probabilities are efficiently derived using the approach described in Bertoldi et al. (2001) by interpolating the frequencies of trigrams of the background (i.e. non adapted) LM with the corresponding frequencies computed on the adaptation text.

The most frequent 128Kwords of the corpus are retained; all the words of the adaptation set are then included in the corresponding lexicon.

## 4 Description of SmarTerp multilingual benchmark

As mentioned above, in SmarTerp we prepared benchmarks for the 3 languages of the project: English, Italian, Spanish. For each language, a number of internet videos having Creative Commons licence were selected, in order to reach at least 3 hours of material on a particular topic, dentistry. Table 3 reports duration and number of words of the benchmarks. Data were collected, automatically transcribed and manually corrected[2] using Transcriber[3], a tool for segmenting, labelling and transcribing speech. In addition to time markers and orthographic transcription of the audio data, we decided to label with parenthesis Important Words (IWs), which represent content words that are significant for the selected domain (i.e. dentistry) and are a fundamental part of the desired output of the automatic system. As only one annotator labelled IWs, it was not possible to compute annotators' agreement for this task. We will address this issue in future works.

| language | recordings | raw duration | transcribed duration | running words | running IWs |
|---|---|---|---|---|---|
| English | 5 | 04:02:34 | 03:03:06 | 28279 | 3343 |
| Italian | 33 | 05:29:34 | 04:10:31 | 31001 | 4560 |
| Spanish | 13 | 03:09:53 | 03:01:59 | 25339 | 3351 |

Table 3: Benchmarks collected and annotated in SmarTerp.

Figure 1 shows a screenshot of Transcriber, where some IWs are highlighted: (dentistry), (dental caries), (periodontal diseases), (oral cancer). In the benchmarks, phrases composed up to 6 words were identified as IWs.

[2]We are really grateful to Susana Rodríguez, who did the manual check for all the languages.

[3]http://trans.sourceforge.net/

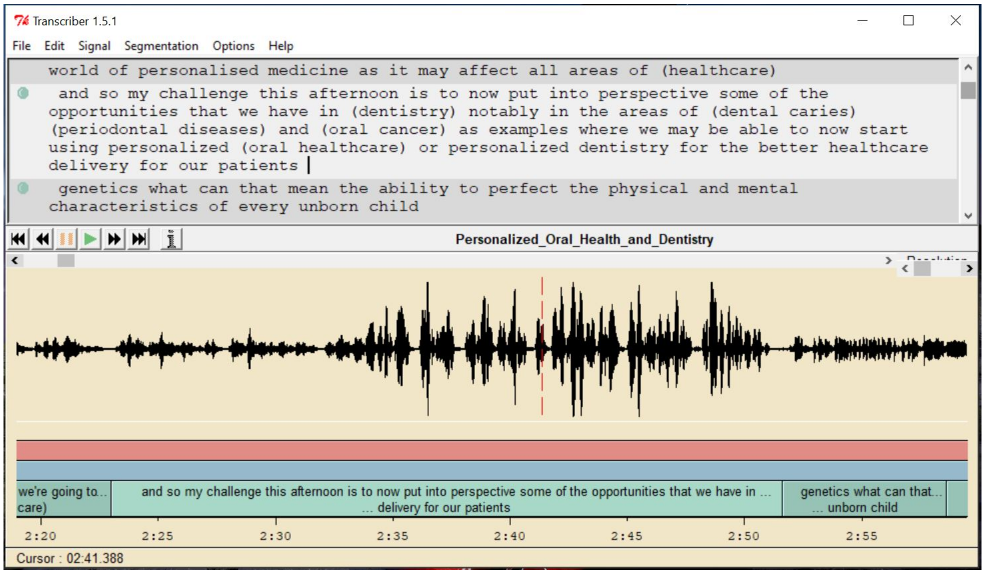


Figure 1: Screenshot of Transcriber, a tool used to manually transcribe the SmarTerp benchmark. In the highlighted segment, IWs are in parentheses.

### 4.1 IW normalization

In order to be able to consistently evaluate the performance of the system in terms of IWs, and considering that it was impossible to pre-define a fixed set of IW patterns, we decided to implement a procedure that automatically processed the whole benchmark. It consisted of the following basic steps, applied independently for every language:

1. identification of all manually defined IWs in the benchmark;
2. reduction to a minimum set of IWs, by removing ambiguities. Given that A, B, C, etc. are single words, some cases are:
   - if exist (A), (B) and (A B), then the IW (A B) is removed - will be replaced by (A) (B);
   - if exist (C), (D E) and (C D E), then the IW (C D E) is removed;
   - note however that if exist (C), (D E) and (D C E), nothing can be removed.
3. regeneration of the benchmark, by applying the following steps:
   (a) remove all round brackets;
   (b) considering the minimum set of IWs, apply new brackets at every IW occurrence, starting from the longest IWs and ending with the one-word IWs;
   (c) in order to evaluate Precision, Recall and F-measure of IWs, remove all words not inside brackets.

Note that some IWs originally present in the benchmark, although legitimate, could not appear in the final version of the benchmark: suppose that the only occurrence of (B) alone is in the context A (B) and also the IW (A B) exist: after the regeneration of the benchmark, both cases will result (A B).

After the application of this algorithm, a consistent version of the benchmark was obtained. By applying the same regeneration steps to the ASR output, a fair comparison was

| | |
|---|---|
| REF<br>ASR<br>ALIGNMENT<br>WER | the most of **them** referred from (pulmonary specialist) **(ENTs)** (paediatricians) **let's let Boyd try** nothing **else**<br>**in** the most of **my** referred from (pulmonary specialist) **ian** (paediatricians) **was led by tried** nothing<br>I_in S_them_my S_ENTs_ian S_let's_was S_let_led S_Boyd_by S_try_tried D_else (Sub= 6 Ins= 1 Del= 1 REF=16)<br>50.00% [ 100 * (6 +1 +1) / 16 ] |
| IW-REF<br>IW-ASR<br>P / R / F | (pulmonary_specialist) **(ENTs)** (paediatricians)<br>(pulmonary_specialist) (paediatricians)<br>Precision 1.00 [ 2 / 2 ] / Recall 0.67 [ 2 / 3 ] / F-Measure 0.80 |
| Isol-IW-REF<br>Isol-IW-ASR<br>P / R / F | (pulmonary) (specialist) **(ENTs)** (paediatricians)<br>(pulmonary) (specialist) (paediatricians)<br>Precision 1.00 [ 3 / 3 ] / Recall 0.75 [ 3 / 4 ] / F-Measure 0.86 |

Table 4: Evaluation metrics on a sample of the English benchmark: WER over the whole text; Precision, Recall, F-measure over both the IWs and the Isolated-IWs. ASR errors are highlighted in bold. IWs are those in parentheses.

possible, considering only the IWs. We could also consider different metrics, either by considering each IW as a single item (despite the number of words that compose it) or by considering separately each word that compose the IWs (henceforth Isol-IW). Standard evaluation of ASR output is Word Error Rate (WER), resulting from a word-by-word alignment between reference text (REF) and ASR output (TEST). In detail, WER is the percentage of substitution, insertions and deletions over the number of REF words. In SmarTerp, however, it could be more useful to concentrate on the IWs only, and to consider Precision, Recall and F-Measure as primary metric. The example in Table 4 shows the different metrics used in this work.

### 4.2 Preliminary analysis

Figure 2 reports OOV rate of the SmarTerp Benchmark for different values of the lexicon size, computed on all the available text data described in Table 2. An inspection of OOV words was

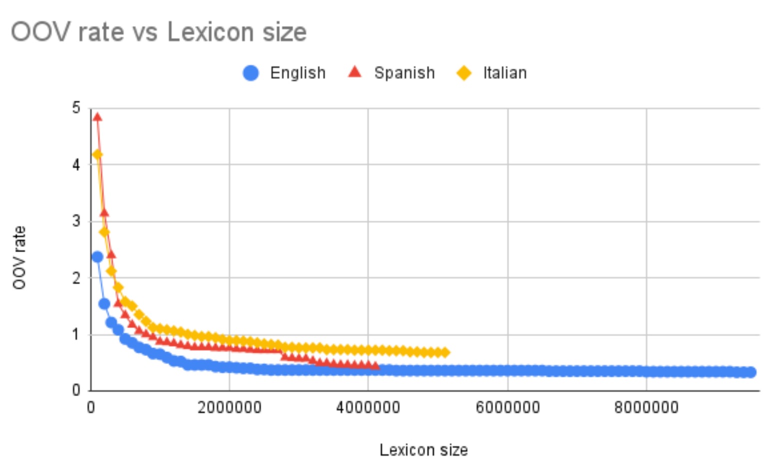


Figure 2: OOV rate of the SmarTerp benchmarks against lexicon size for the 3 languages.

| allunghiamo *we lengthen* | | distinguerle *distinguish them* | | divideremo *we will divide* | |
|---|---|---|---|---|---|
| 10355 | allunga | 12118 | distingue | 7273 | divide |
| 12657 | allungare | 12493 | distinguere | 7931 | dividendo |
| 17187 | allungato | 20484 | distinguono | 12286 | dividere |
| 18040 | allungo | 26323 | distinguo | 14127 | dividendi |
| 20126 | allungamento | 34366 | distinguersi | 15601 | dividono |
| 23870 | allungano | 52496 | distinguendosi | 27370 | dividersi |
| 25749 | allungata | 56673 | distingueva | 43165 | divideva |
| 35514 | allungando | 60858 | distinguerlo | 59956 | dividerà |
| 40996 | allungate | 61213 | distinguendo | 61370 | dividerci |
| 42540 | allungati | 67741 | distinguibili | 62319 | divideranno |
| 43104 | allungarsi | 75608 | distinguerla | 63369 | dividendosi |
| 60394 | allunghi | 77105 | distinguibile | 68113 | dividevano |
| 98044 | allungherà | 79891 | distinguevano | 80977 | dividerli |
| 106019 | allungava | 91152 | distinguerli | 84294 | dividend |
| 120007 | allungandosi | 115236 | distinguiamo | 91609 | divida |
| 126079 | allungherebbe | 116550 | distingua | 97706 | dividiamo |
| | | 119097 | distinguerà | 121708 | dividerlo |

Table 5: Morphological variations of OOV words, known in the 128 Kwords lexicon, along with their position in the lexicon.

done for the Italian language, in order to better understand how the OOV words are distributed among different classes. With respect to the 128 Kwords lexicon, we had that the Italian benchmark is composed of 31001 running words, of which 1089 are OOV (corresponding to 3.51% OOV rate). The number of different OOV words was 474, manually classified as follows:

- **190 Morpho**: morphological variations of common words (e.g. allunghiamo, distinguerle, divideremo - *we lengthen, distinguish them, we will divide*);
- **181 Tech**: technical terms, that will be part of IWs so it is extremely important to keep their number as low as possible (e.g. bruxismo, implantologia, parodontopatici - *bruxism, implantology, periodontal disease*);
- **34 Errors**: words that should not be here and will be fixed soon: numbers in letters, wrong tokenization (e.g. cinque, computer-assistita, impianto-protesica, l'igiene - *five, computer-assisted, implant-prosthetic, the hygiene*);
- **28 English**: terms in English, often they are technical terms and should be recognized (e.g. osteotomy, picking, restaurative, tracing);
- **20 Names**: proper names of people, firms or products (e.g. claronav, davinci, hounsfield, navident);
- **10 Latin**: latin words (e.g. dolor, restitutio, tumor - *pain, restoration, swelling*);
- **8 Acronyms**: (e.g. t-test, mua, d3, d4);
- **3 Foreign**: pseudo-foreign words that need particular care for pronunciation (e.g. customizzata, customizzati, matchare - *Italian neologisms from English custom, match*).

Tech, English, Names, Latin and Foreign will deserve a particular attention in future studies, because they are important for the domain. Errors will be fixed and should disappear; Acronyms should be recognized as subwords (e.g., d3 as d 3). Morpho will probably be misrecognized as another morphological variation of the same stem, present in the active dictionary, which in this domain is not considered a critical error. Note that a single verbal stem in Italian can generate up to 300 different words in Italian, including clitics. In Table 5 you can see the morphological variations of the 3 terms of the class Morpho reported above which are present in the 128 Kwords lexicon.

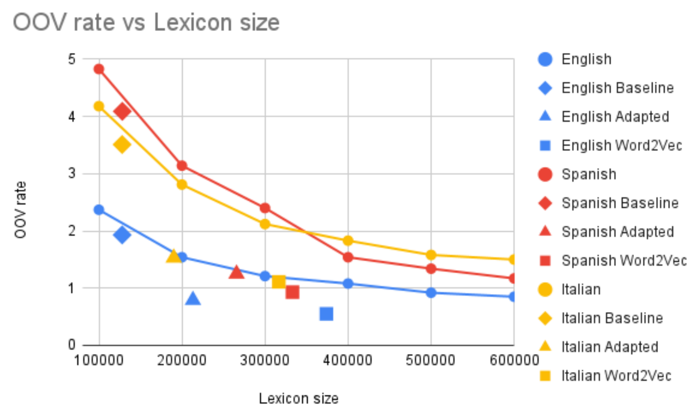


Figure 3: OOV rate of the SmarTerp benchmarks against lexicon size for the 3 languages, for all the experiments and languages.

## 5 Experiments and results

Since several approaches can be employed to obtain, enlarge and use the seed words (e.g. based on texts distance, texts semantic similarity, etc) we consider the following indicators that allow to measure their effectiveness on the benchmarks collected and manually transcribed within the SmarTerp project.

- Seeds: number of seed words, used to extract the adaptation text:
- Out Of Vocabulary rate (OOV rate): it is the percentage of unknown words in the benchmark, with respect to the lexicon. OOV words cannot be part of the output of the ASR, hence they will be certainly errors. We should try to get a low OOV rate without the lexicon size growing too much;
- Lexicon size: total number of active words in the adapted LM;
- Word Error Rate (WER): it measures the percentage of errors made by the ASR;
- Precision, Recall, F-Measure over the set of Important Words (IWs) that were defined.

The following experiments were carried out for each of the three languages:

- **Baseline**: the initial 128Kwords lexicon and the LM trained on the whole corpus, without any adaptation;
- **Adapted**: LM adapted starting from seed words coming from a dental glossary (normally 2-3 pages of text, resulting into some hundreds of seeds), found with a quick search in internet for terms like "dental glossary" (e.g. https://bnblab.com/intro/terminology).
- **Word2Vec**: LM adapted using seed words obtained from 5 initial seed words, applying two iterations ($I_w = 2$) of the procedure based on semantic similarity and retaining, for each term, $N_w = 40$ words, obtaining $\sim 3000$ seed words. The 5 magic words[4] were:
  - **English**: tartar, filling, caries, tooth, dentist

[4]Many thanks to Susana Rodríguez for the translations of the magic words from Italian

- **Italian**: tartaro, otturazione, carie, dente, dentista
- **Spanish**: sarro, relleno, caries, diente, dentista

Figure 3 reports OOV rate of the SmarTerp benchmark for different values of the lexicon size for each experiment, along with the initial part of the curve of Figure 2. It should be noted that, for every language, Baseline is along the initial curve, while both Adapted and Word2Vec are well below it. For all languages, Adapted has a Lexicon size which is in between Baseline and Word2Vec. This is due to an initial choice of the parameters described in Section 2: by changing the parameters, a cloud of values could be generated instead of a single point. In fact, in this work we report only initial experiments and future efforts will be devoted to a parameter optimization. In any case, the Lexicon size is directly related to the number of seeds and on the size of the adaptation text, which plays a very important role in the adaptation stage.

Table 6 reports preliminary results on the three benchmarks, for all the experiments. Together with the number of obtained seed words, OOV rate and Lexicon size, we report WER computed on all the uttered words (including functional words, which are useless for this task), and Precision/Recall/F-measure computed both on IWs and Isol-IWs: since they represent the most technically significant words in the domain, they are more related to the output desired by interpreters. It is worth noting that, with respect to Baseline, both the Adapted and Word2Vec systems are effective for all of the three languages and for all the considered metrics. Word2Vec performs slightly better than Adapted, but this can be due to the initial value of the parameters that bring to more seeds and to a bigger Lexicon size. Low WER for English is partly due to a scarce audio quality in the recordings, that mainly affects functional words: this explains the English high precision, which is computed on IWs only.

| | Seeds | Lex size | OOVrate | WER | IW P / R / F | Isol-IW P / R / F |
|---|---|---|---|---|---|---|
| Eng BL | 0 | 128041 | 1.93% | 26.39% | 0.90 / 0.61 / 0.73 | 0.96 / 0.59 / 0.73 |
| Eng ada | 257 | 213237 | 0.79% | 23.34% | 0.92 / 0.73 / 0.81 | 0.97 / 0.71 / 0.82 |
| Eng w2v | 2999 | 373956 | 0.55% | 23.86% | 0.93 / 0.72 / 0.81 | 0.97 / 0.70 / 0.81 |
| Ita BL | 0 | 128009 | 3.51% | 15.14% | 0.88 / 0.67 / 0.76 | 0.95 / 0.67 / 0.79 |
| Ita ada | 213 | 190126 | 1.53% | 11.73% | 0.96 / 0.84 / 0.89 | 0.98 / 0.82 / 0.90 |
| Ita w2v | 3527 | 316679 | 1.11% | 11.28% | 0.96 / 0.85 / 0.90 | 0.99 / 0.84 / 0.91 |
| Spa BL | 0 | 128229 | 4.09% | 22.60% | 0.86 / 0.56 / 0.68 | 0.93 / 0.56 / 0.69 |
| Spa ada | 673 | 265764 | 1.25% | 17.74% | 0.95 / 0.76 / 0.85 | 0.98 / 0.75 / 0.85 |
| Spa w2v | 3207 | 333072 | 0.93% | 17.31% | 0.95 / 0.79 / 0.86 | 0.98 / 0.78 / 0.87 |

Table 6: Preliminary results for Baseline (BL), Adapted (ada) and Word2Vec (w2v) systems. Both WER on all words and Precision/Recall/F-measure on composite and isolated IWs are reported.

## 6 Conclusions

We described two different approaches for extending the dictionary of an ASR system in order to detect important terms from technical speeches, namely dental reports, to be translated by simultaneous professional interpreters. The two approaches consist in extracting adaptation text from a huge set of text data, starting from some seed words. In the first one, seed words come from a given glossary. The second one is based on the application of a text similarity measure to an initial (very small) set of 5 seed words. After the application of the selection procedures we adapted the language models used in the ASR system employed in a computer assisted interpretation (CAI) system under development and we proved the effectiveness on the approaches in terms of different evaluation metrics.

## References


Amin, K., Lancaster, G., Kapetanakis, S., Althoff, K., Dengel, A., and Petridis, M. (2019). Advanced similarity measures using word embeddings and siamese networks in cbr. In *Proc. of IntelliSys*, volume 1038.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Bell, P., Fainberg, J., Klejch, O., Li, J., Renals, S., and Swietojanski, P. (2021). Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

Berg, A., O'Connor, M., and Cruz, M. T. (2021). Keyword transformer: A self-attention model for keyword spotting.

Bertoldi, N., Brugnara, F., Cettolo, M., Federico, M., and Giuliani, D. (2001). From broadcast news to spontaneous dialogue transcription: Portability issues. In *Proc. of ICASSP*, Salt Lake City, UT(US).

Bridle, J. (1973). An efficient elastic-template method for detecting given words in running speech. In *British acoustical society spring meeting*, pages 1—-4.

Chen, G., Parada, C., and Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *Proc. of ICASSP*, page 4087–4091.

Desmet, B., Vandierendonck, M., and Defrancq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In *Interpreting and technology*, pages 13—27, C. Fantinuoli ed. Berlin: Language Science Press.

Fernandez, S., Graves, A., and Schmidhuber, J. (2007). An application of recurrent neural networks to discriminative keyword spotting. In *Artificial Neural Networks – ICANN 2007*, page 220–229.

Gretter, R. (2014). Euronews: a multilingual speech corpus for ASR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2635–2638, Reykjavik, Iceland. European Language Resources Association (ELRA).

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proc. of International Conference on Machine Learning*, Beijing, China.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proc. of NIPS*, volume 2.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. IEEE Catalog No.: CFP11SRW-USB.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. of INTERSPEECH*, pages 2751–2755.

Rodríguez, S., Gretter, R., Matassoni, M., Falavigna, D., Alonso, Á., Corcho, O., and Rico, M. (2021). SmarTerp: A CAI system to support simultaneous interpreters in real-time. In *Proc. of TRITON 2021*.

Rose, R. and Paul, D. (1990). A hidden markov model based keyword recognition system. In *Proc. of ICASSP*, page 129–132.

Sainath, T. and Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. In *Proc. of Interspeech*.

Song, Y., Jiang, D., Zhao, W., Xu, Q., Wong, R. C.-W., and Yang, Q. (2019). Chameleon: A language model adaptation toolkit for automatic speech recognition of conversational speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 37–42, Hong Kong, China. Association for Computational Linguistics.

Vijaymeena, M. and Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(1).

Weintraub, M. (1995). Lvscr log-likelihood ratio scoring for keyword spotting. In *Proc. of ICASSP*, volume 1, page 297–300.

Yan, H., He, Q., and Xie, W. (2020). Crnn-ctc based mandarin keywords spotting. In *Proc. of ICASSP 2020*, pages 7489–93.