

Standardizing Language with Word Embeddings and Language Modeling in Reports of Near Misses in Seveso Industries

Simone Bruno*, Silvia Maria Ansaldi⁺, Patrizia Agnello⁺, Fabio Massimo Zanzotto*

* University of Rome Tor Vergata, fabio.massimo.zanzotto@uniroma2.it

⁺ INAIL, {s.ansaldi,p.agnello}@inail.it

Abstract

Standardizing technical language has always been a strong necessity of the technological society. Today, Natural Language Processing as well as the widespread use of computerized document writing can give a tremendous boost in reaching the goal of standardizing technical language. In this paper, we propose two methods for standardizing language. These methods have been applied to the dataset of near misses, collected during the inspections at Major-Accident Hazard (MAH) Industries.¹

1 Introduction

Standardizing technical language has always been a strong necessity of the technological society. Artifacts, objects, measures and so on should have a clear name and a clear description in order to assure mutual understanding, which leads to the reach of important goals in building and controlling machines. However, language standardization has always the same problem: language is a *social phenomenon* (de Saussure, 1916). Hence, whenever a group gather for designing or using a technical object, this group can develop a specific sub-language or just adapt the shared technical language. This adapted sub-language can be then effectively used to refer to parts of this technical object. It is sufficient that group members agree upon this language and the mutual understanding occur. Yet, the language used by the specific group may prevent the others to understand what is written.

Nowadays, Natural Language Processing as well as the widespread use of computerized document writing can give a tremendous boost in reaching the goal of standardizing technical language. Language in use can be captured and, then,

analyzed. Technical people can be invited to use a standardized dictionary with writing suggestions.

This paper discusses two different methods of standardizing technical languages, which have been applied to a dataset of near misses coming from the inspections at Major-Accident Hazard (MAH) industries, named also “Seveso” industries. . The first method aims to help a standardization agency to propose the standard language for writing these reports. We proposed to analyze language in use by word embedding similarity such that the standardization agency can propose a language that is close to the one used. The second method aims to reduce the use of unnecessary synonyms in compiling reports of near misses. In fact, using unnecessary synonyms may result in confusing the report. For this problem, we propose to use a combination of language modeling derived from the CBOW model of the word2vec (Mikolov et al., 2013) along with a classical cosine similarity using word embeddings. We experimented with a dataset of anonymized reports of near misses from Seveso Industries, which INAIL has institutionally collected.

The rest of the paper is organized as follows. Section 2 describes the application scenario and the dataset. Section 3 shortly reports on the models used in this study and proposes the two tasks. Section 4 reports on a preliminary analysis of the possible results of the system. Finally, Section 5 draws some conclusions and proposes further investigations.

2 Background

2.1 Scenario

The European “Seveso” Directive deals with the control of major-accident hazards involving dangerous substances, which can cause toxic clouds, fire, or explosion with consequences to people, as-

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Ref: 66	Data (Date): 2007-02-15	
Titolo (Title): Descrizione (Description):	Trasudamento OCD da serbatoio di stoccaggio OCD Durante le operazioni di riempimento del serbatoio K2 da nave cisterna, si è notato un leggero trasudamento diOCD per corrosione del mantello (sottospessore localizzato mantello serbatoio) a quota 6 metri circa lungo il latoovest. Uno degli operatori addetto ai controlli durante la discarica della nave ha evidenziato l'evento. L'operazionedi discarica della nave cisterna è stata fermata. Non si sono avuti rilasci, a meno del leggero trasudamento.	
Sistemi tecnici critici (Critical Technical System):	serbatoio	
Sostanza (Substance):	olio combustibile (ocd)	
Fattori gestionali (Managing Factor)	Descrizione (Description)	Azioni pianificate (Planned Actions)
4.iv	Fallimento procedure di manutenzione e controllo.	Fuori servizio e bonifica del serbatoio.

Figure 1: Sample Report of a Near Miss within the European Seveso Directive - Italian Localization: Translation is provided for Field Names

sets and environment, also outside the establishments. All European Member States apply this Directive, which foresees periodical inspections by National Competent Authorities; in Italy, Inail is one of these authorities. During the inspection, the operator has to provide the inspectors with the list of near-misses, minor incidents, and accidents occurred in the last ten years. Near misses and minor incidents are events of losses of containment, involving dangerous substances with none or minor consequences, respectively. In Seveso industries, the registration and the analysis of near misses is strongly recommended, as they can be considered as precursors of incidents with serious consequences.

In Italy, under Seveso legislation, there are about a thousand industries, including refineries, petrochemical, and chemical. One of the pillars of the Seveso Directive is the Safety Management System SMS, whose adoption is mandatory for the establishments' operators, in order to control major accident hazards.

The Safety Management System (SMS), implemented by the establishment's operator, addresses technical measures and organizational procedures in order to guarantee human, asset and environmental safety, with a view to the prevention of major accident or the mitigation of their consequences.

In the recent inspections, the focus is often toward the study of the incidents and near misses (see Figure 1). The approach based on near-miss discussion is considered more "risk based" as it is able to single out the critical issues of the safety system.

2.2 Corpus

The dataset refers to the near misses reports provided by the operators of "Seveso" establishments. The collection of reports on near misses, hereafter referred as REP corpus, consists of 1300 documents called "operative experiences". These operative experiences span the period from 2006 to 2017 and are related to 320 plants.

Each "operative experience" tells about the events occurred in the recent past (see Figure 1 for an example). Each event is registered by the operators filling in a pre-defined form. The document contains information including the date, a title summarizing the event, a short description, the reference to failed, missing or misapplied technical or procedural barriers, those that stopped the escalation and the recovering actions, and eventually the planned actions for improving the safety.

It is out of scope of this paper to discuss the different methods used in the literature to manage near miss information for improving the safety management system. However, the common objective is to exploit the valuable information contained. (Ansaldi et al., 2018) describe a method to extract knowledge from this collection of documents, and to support foresights or intuitions about the safety of process industries. Another application has been developed for understanding if the lessons from major accidents have been fully learnt and implemented (Ansaldi et al., 2016). The issue has been addressed by looking for similarities between near misses and accident characteristics, and by evaluating their semantic distance.

Although the form of the document is the same adopted for all operators, the compiling mode

varies by the establishments and by the type of event recorded. The accuracy of the documents is not homogeneous and the interpretation of operative experience concept changes from one establishment to another; their carefulness varies on the sector activities, and often reveals the safety culture of the establishment. At a few establishments, just the releases of hazardous substance without consequences are registered. In other cases, reports include anomalies, unsafe situations, failures, and trivial errors; that is, events not directly related to major accident hazard. The documents are various, but represent truthful pictures of deviations occurred inside the establishment.

3 Methods

The overall goal is to show that existing methodologies can help in standardizing language in the specific case of reports on near misses on Seveso industries and we aim to perform this standardization with two tools: (1) analyzing similarities among words in current reports; (2) propose a methodology to help in writing these reports.

3.1 Challenges

The specific case of reports on near misses is particular for several compelling reasons. The first compelling reason is that reports are written by operators belonging to sub-communities of speakers. In fact, people working in each plant can be considered a sub-community, which shares a particular language. Hence, standardizing language of reports means also harmonize sub-languages of different sub-communities, which do not interact. This problem is particularly severe when the aim is to standardize language across the whole Seveso industries. The second compelling reason is the different background of reports' writers. Reports are in fact written by operators, which may have different knowledge, different school degree, and different cultural background. This reason makes particularly relevant the goal to help writers in compiling reports on near misses.

3.2 Enabling Tools and Methodologies

To meet the overall goal, we here experiment with standard and well-assessed models and methodologies: the notion of word embedding. In fact, the long tradition of representing word meaning in vectors is what is needed to: (1) help the standardization organism to develop a common and accept-

able language; (2) devise ways to suggest more appropriate words to writers of reports. In this study, we used two different word embeddings:

- General Language Word Embeddings (GLwe)(Cimino et al., 2018): these are word embeddings pre-trained with word2vec (Mikolov et al., 2013) on a general purpose corpus of the Italian language, that is, itWaC (Baroni et al., 2009)
- Domain-adapted Word Embeddings (Dawe): these are word embeddings obtained training word2vec (Mikolov et al., 2013) using GLwe as initialization and the REP training corpus

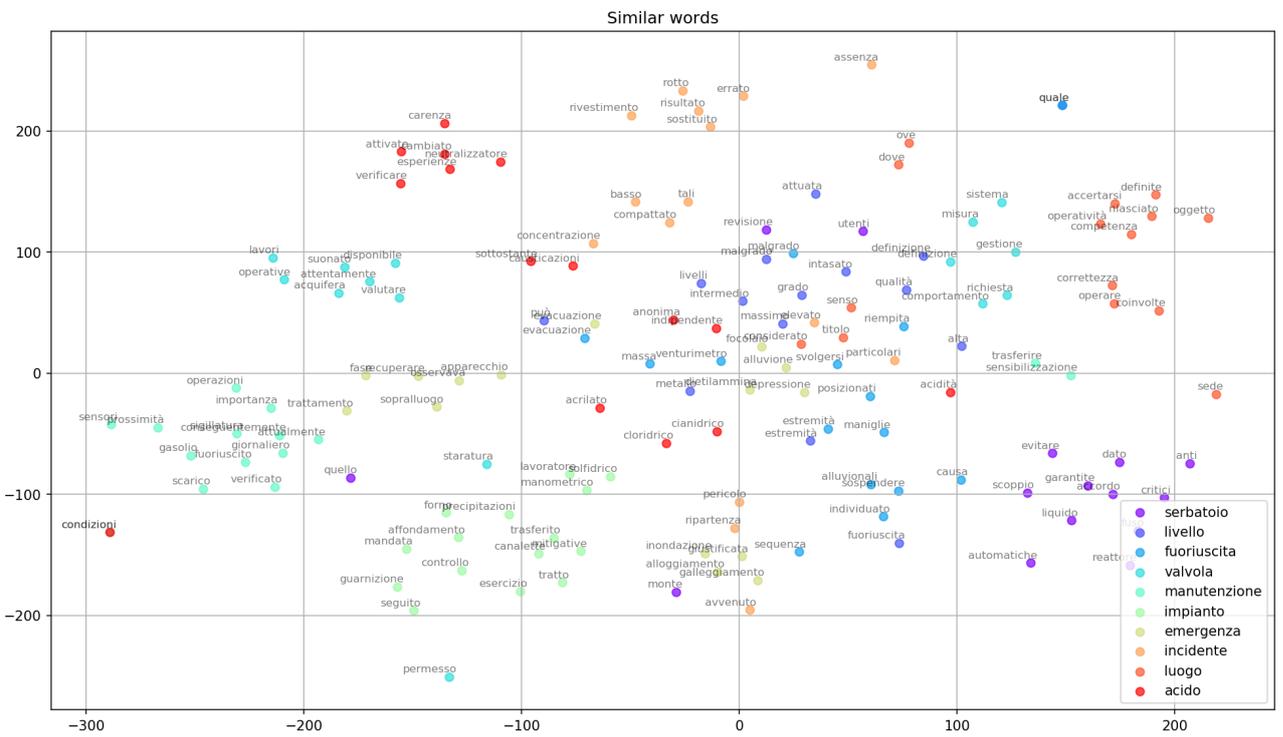
3.3 Task 1: Understanding Language of Near-Miss Reports

We aim to provide the standardization organism, that is, INAIL, the possibility to investigate the language used in these reports on near misses. The possibility we explored is to provide a visual representation of similarity computed using similarity among word embeddings. Giving this visual representation, researchers in INAIL can devise the definition of a standard language that is built on a common and shared language. This idea is similar to what has been done in the past for terminology extraction. The real added value is that similarity among terms is computed according to word embeddings.

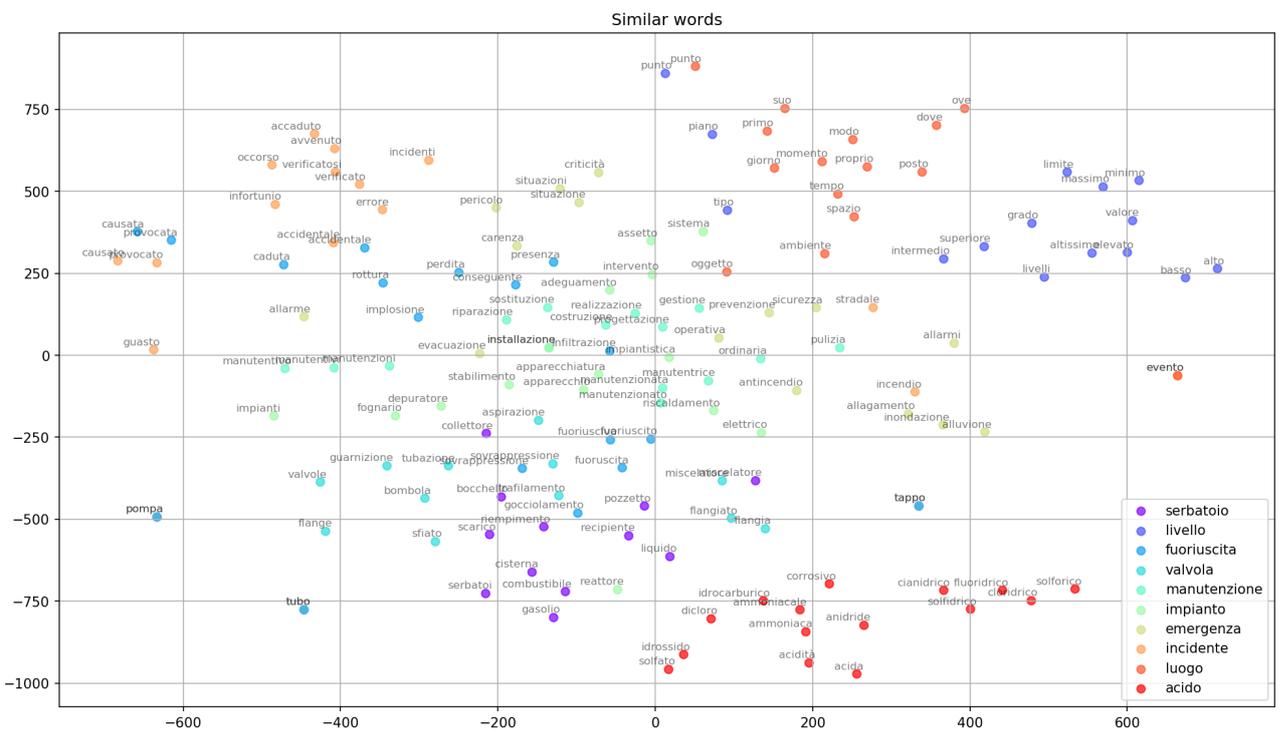
3.4 Task 2: Standardizing Report with Assisted Writing

We aim to provide a tool to assist operators while writing reports. We explored the first capability of this tool, that is, avoiding unnecessary use of synonyms while writing. In the Italian tradition, using repeating words is seen as bad writing. Hence, when writing, synonyms are used to introduce a variation. However, for technical documents, unnecessary use of synonyms in core concepts may introduce misunderstanding. Hence, we envisage a tool that helps in reducing use of synonyms.

The algorithm governing the tool works as follows. While writing a report, the algorithm accumulate words in a set W . Whenever a new content word w is added, the algorithms compute the similarity with the words in the set W . If there is a word $w' \in W$ for which the similarity $sim(w, w') = w^T w'$ is above a threshold τ , the algorithm suggests w' as a possible substitution of w . In this way, the operator is forced to



(a) General Language Word Embeddings



(b) Domain-adapted Word Embeddings

Figure 2: Similarities among Words: Studying and Understanding Technical Language with Word Embeddings

#	Text
174	La perdita non si era evidenziata al controllo dell'area effettuato preliminarmente all'inizio attività, né rilevata dal CTM presente in <i>zona</i> <u>area</u> (<i>sim</i> = 0.64) <u>attività</u> (<i>sim</i> = 0.31)
175	... Alle ore 10,15 il CT rilevava visivamente la presenza di tracce di virgin nafta miscelati con le acque di scarico e, mentre si accingeva a chiudere la valvola sul dreno di fondo colonna, improvvisamente, si sviluppava un principio d'incendio. Lo stesso CT, utilizzando le manichette di erogazione acqua già attive per il lavaggio dell'area atto a favorire il convogliamento dei reflui nel pozzetto di raccolta di raffineria, estingueva prontamente il <i>focolaio</i> <u>incendio</u> (<i>sim</i> = 0.46) <u>intervento</u> (<i>sim</i> = 0.34)
109	Necessità di prevedere un più elevato grado di protezione contro la perdita di contenimento da fondo serbatoi. La <i>fuoriuscita</i> <u>perdita</u> (<i>sim</i> = 0.54) <u>contenimento</u> (<i>sim</i> = 0.42)

Figure 3: Suggested replacements for with already used synonyms

think whether the word w' that *s/he* already used is similar to the word *s/he* is using now. In this case, w' can be used to replace w and an unnecessary synonym is avoided.

4 Experimental Results

4.1 Task 1

For the first task, we experimented with the two dictionaries: the General Language word embeddings (GLwe) and the Domain-adpated word embeddings (Dawe). Similarity spaces for the two word embeddings (see Figure 2) may help in understanding whether unnecessary synonyms are used and, hence, suggest a standardized word that should be used for a group of words.

Using the two dictionaries, we built two similarity spaces (Figure 2) obtained as follows. We selected 10 frequent words in the REP training corpus and, then, we presented in the two figures the top 15 words that are more similar to the 10 selected frequent words. The similarity spaces are built according to GLwe (Figure 2a) and according to Dawe (Figure 2b).

The Dawe similarity space (Figure 2b) gives apparently better hints on how words are used. The dictionary seems to be more tailored to the specific domain. In fact, there is an interesting groups of words such as $\{avvenuto, accaduto, occorso, verificatosi\}$ and $\{causato, provocato\}$. These groups are missing in the GLwe similarity space (Figure 2a).

4.2 Task 2

For the second task, we experimented with some sample reports. The algorithm in action is reported in Figure 3. This test has been carried out on existing reports and aimed to show that some words can be replaced with previously used words. In the report #174, the word *zona* can be replaced with the word *area*, which has been previously used. In the report #175, the word *focolaio* could be replaced with the word *incendio*. Finally, in the report #109, the word *fuoriuscita* can be replaced with the word *perdita*. However, the operator is free to accept or refuse the suggestion if this is not satisfactory.

5 Conclusion and Future Work

Standardizing language is a need of our technological society. In this paper, we investigated the possibility of using modern NLP techniques to reach this goal in the specific scenario of near misses in Seveso Industries. Initial results on the corpus provided by Inail are interesting and leave room for improvement. Future model should include the treatment of multi-word expressions by using compositional distributional semantic models (Zesch et al., 2013; Zanzotto et al., 2015), should merge distributional and ontological models, and should include a clear model for repaying knowledge producers (Zanzotto, 2019).

References

- Silvia Maria Ansaldi, Patrizia Agnello, and Paolo Bragatto. 2016. [Incidents triggered by failures of level sensors](#). *Chemical Engineering Transactions*, 53:223–228.
- Silvia Maria Ansaldi, Annalisa Pirone, Rosaria Vallerotonda Maria, Paolo Bragatto, Patrizia Agnello, and Corrado Delle Site. 2018. [How inspections outcomes may improve the foresight of operators and regulators in seveso industries](#). *Chemical Engineering Transactions*, 67:367–372.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. [The wacky wide web: a collection of very large linguistically processed web-crawled corpora](#). *Language Resources and Evaluation*, 43(3):209–226.
- Andrea Cimino, Lorenzo De Mattei, and Felice Dell’Orletta. 2018. [Multi-task learning in deep neural networks at EVALITA 2018](#). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *CoRR*, abs/1301.3781.
- Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.
- Fabio Massimo Zanzotto. 2019. [Viewpoint: Human-in-the-loop Artificial Intelligence](#). *Journal of Artificial Intelligence Research*, 64:243–252.
- Fabio Massimo Zanzotto, Lorenzo Ferrone, and Marco Baroni. 2015. [When the whole is not greater than the combination of its parts: A ”decompositional” look at compositional distributional semantics](#). *Comput. Linguist.*, 41(1):165–173.
- T. Zesch, I. Korkontzelos, F.M. Zanzotto, and C. Biemann. 2013. [Semeval-2013 task 5: Evaluating phrasal semantics](#). volume 2, pages 39–47. Cited By 12.