Las descripciones lingüísticas de datos en los sistemas «Data to Text»

A. Ramos-Soto, A. Bugarín, and S. Barro

Centro de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela
{alejandro.ramos, alberto.bugarin.diz, senen.barro}@usc.es

Resumen En este trabajo se realiza una revisión del estado del arte de los modelos y sistemas de generación de narrativas descriptivas en lenguaje natural a partir de datos y se analiza el papel que juegan las aproximaciones basadas en computación flexible dentro de este contexto. Estas aproximaciones (denominadas descripciones lingüísticas de datos) se han ido desarrollando, en general, en paralelo con el ámbito más general de los modelos de Generación de Lenguaje Natural y en concreto los sistemas 'Data to Text'. Describimos los modelos, aplicaciones y métodos de evaluación más relevantes de los ámbitos, así como una valoración sobre sus actuales y potenciales puntos de convergencia.

Keywords: Descripciones lingüísticas de datos, Generación de Lenguaje Natural, Sistemas 'Data to text'

1. Introducción

La tarea de generar información expresada en lenguaje natural ha venido siendo abordada tradicionalmente, de forma independiente, por dos campos de investigación: el ámbito de la Generación de Lenguaje Natural (NLG) y el de la Descripción Lingüística de datos (LDD). NLG y más específicamente el subcampo 'Data to text' (D2T) se ha enfocado, por mas de treinta años, a la generación, a partir de datos, de narrativas o textos lo mas difícilmente distinguibles de los que crearía un ser humano. La complejidad de los sistemas desarrollados en este ámbito es muy notable y se han establecido diversas técnicas y metodologías (no únicas) que guían la construcción de las soluciones. La gran cantidad de sistemas NLG existentes en la actualidad (al menos 400 [1], aunque es posible intuir que pueden ser mas) es una evidencia del interés e impacto que ha obtenido este campo de investigación. Por su parte, la LDD, que se origina en el ámbito de la computación flexible, proporciona sumarios o descripciones que involucran términos imprecisos utilizando conjuntos, particiones u operadores borrosos. Se trata de un campo de investigación más reciente, que proporciona información -esencialmente- en forma de términos lingüísticos. Su desarrollo mas intenso se dio en los últimos veinte años, con los avances en el campo de la computación con palabras y percepciones [2,3,4]. La mayoría de aproximaciones en este ámbito

han sido teóricas, aunque en algunos casos se han planteado ejemplos o problemas de tipo práctico. Sin embargo, pese a este desarrollo paralelo de D2T y LDD, están comenzando a surgir aproximaciones híbridas que combinan ambas aproximaciones para proporcionar soluciones a problemas reales.

2. Sistemas Data-To-Text (D2T)

La Generación de Lenguaje Natural (NLG) es el ámbito de la lingüística computacional que trata acerca del como producir automáticamente textos en lenguaje natural. La demanda de este tipo de sistemas es creciente, especialmente cuando los textos o narrativas se deben construir a partir de datos (habitualmente numéricos) para describir la información más relevante implícita en ellos. Este es el ámbito de los sistemas denominados 'Data To Text' (D2T), de los que podemos encontrar un buen número de aplicaciones de todo tipo en ámbitos como la meteorología, e-salud, inteligencia de negocio, procesos industriales entre otros [1,5,6].

Existen diversas arquitecturas y propuestas para el diseño de los sistemas D2T, que suelen depender del problema a tratar y del desarrollador del mismo. Sin embargo, existe un cierto consenso sobre cuales son las tareas básicas a abordar por este tipo de sistemas. Así, [7],[8] han descrito una serie de tareas genéricas que se deben abordar de forma secuencial. En términos generales, la tarea principal de convertir los datos de entrada en un texto descriptivo de salida se subdivide en las siguientes seis actividades básicas:

1. Determinación de contenidos: proceso en el que se decide qué información se comunicará en la narrativa. Consiste en la creación de un conjunto de mensajes básicos resumidos a partir de los datos de entrada y expresados en algún lenguaje intermedio que distingue etiquetas, entidades, conceptos y relaciones de interés en el dominio de aplicación.
2. Planificación del discurso: proceso por el que se dota de orden y estructura al conjunto de mensajes a transmitir.
3. Agregación de sentencias: proceso que agrupa mensajes en sentencias, lo que mejora en ocasiones la fluidez del texto.
4. Lexicalización: proceso donde se decide que palabras y expresiones específicas se van a usar para expresar los conceptos y relaciones del dominio.
5. Generación de expresiones de referencia: selección de palabras o expresiones que identifican entidades del dominio. Se caracteriza por ser discriminante, asegurando que el sistema proporciona información suficiente para diferenciar cada entidad del resto.
6. Realización lingüística: aplicación de reglas gramaticales para producir un texto final correcto desde el punto de vista sintáctico, morfológico y ortográfico.

Aunque, desde un punto de vista pragmático, la arquitectura habitual se reduce a una secuencia de tres actividades:

1. Planificación de texto (agrupa las etapas 1 y 2)
2. Planificación de las sentencias (agrupa las etapas 3 a 5)
3. Realización lingüística

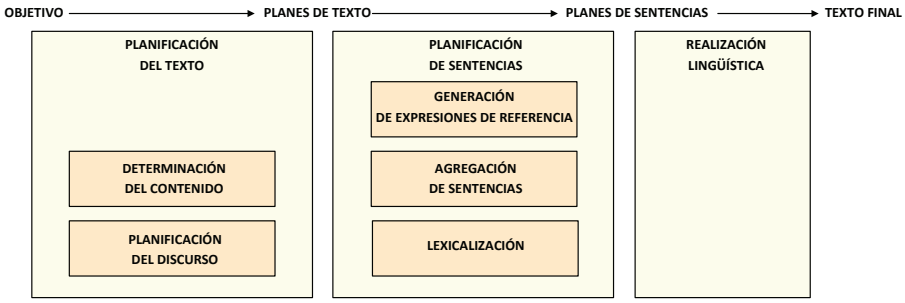

Figura 1. Arquitectura y actividades de un sistema NLG.

Los sistemas D2T se han utilizado con muy distintos propósitos, que vand desde los sistemas de diálogo hasta los diferentes tipos de interacción o retroalimentación de información a los usuarios, y en muy diversos ámbitos de aplicación. Sin ánimo de ser exhaustivos, tenemos, entre otros, los siguientes:

- Meteorología: sistemas pioneros en la generación de pronósticos meteorológicos como FoG [9], MultiMeteo [10] o SumTime-Mousam [11], que proporcionan información multilingüe que, en el último caso, incluye diferentes niveles de detalle en función del perfil de usuario, así como el reciente servicio del British Met Office [12] que genera informes diarios para las poblaciones del Reino Unido. TEMSIS [13] y MARQUIS [14] generan informes medioambientales sobre el estado de la calidad del aire y otras variables de interés.
- Salud: además de SUREGEN-2 [15], que produce documentos médicos relativos a hallazgos clínicos o procedimientos, la familia de sistemas BabyTalk[16] es el sistema quizá de mayor impacto, puesto que genera informes textulaes a partir de los datos fisiológicos de bebés ingresados en Unidades de Cuidados Intensivos Neonatales. La agregación de información heterogénea y de datos de diferentes fuentes, así como los distintos perfiles de usuario que contempla (enfermería, médicos, padres) hacen de Babytalk uno de los sistemas más completos de entre los descritos en la bibliografía.
- Industria: de forma general existen diversos sistemas que ayudan a los usuarios en tareas como la generación de documentos, cartas o la monitorización de procesos. Así, Project Reporter [17] describe el progreso de las tareas de un proyecto, monitorizando el estado de elementos como el personal, gastos o costes. Patent Claim Expert [18] genera borradores para la descripción de patentes siguiendo un modelo preestablecido. SumTime-Turbine [19] es una solución que monitoriza datos de turbinas de gas o de explotaciones

petrolíferas, detecta sobre ellos patrones relevantes y construye informes de análisis o alertas en tiempo real.

Un aspecto relevante en los sistemas D2T es que, dada su complejidad y frecuente su aplicación al mundo real, requieren la aplicación de metodologías de evaluación a través de las cuales se valore en que medida cumplen los requisitos de los expertos. Aun siendo un tema abierto en la investigación en NLG, si hay establecidos algunos consensos a este respecto. En general, la mayoría de métodos de evaluación son cuantitativos [20] y tratan de obtener algún tipo de medida numérica sobre la calidad de su desempeño. Métodos habituales son la realización de cuestionarios a expertos humanos para que evalúen las narrativas generadas o la evaluación mediante métricas de la similaridad entre los textos generados y un corpus [21] representativo. Hay también métodos cualitativos complementarios que suelen utilizarse para identificar y corregir aspectos en las etapas de análisis del contenido y del discurso [22], [23].

De forma general, podemos indicar que los sistemas D2T, aunque generan narrativas que suelen incluir palabras o términos imprecisos, no suelen incluir modelos que manejen explícitamente dicha incertidumbre, como son los que proporciona la computación flexible. Aunque etapas como la determinación del contenido suelen describirse de forma muy sucinta en los trabajos, si puede decirse que con carácter general no hay menciones explícitas relevantes al uso de modelos o técnicas borrosas en la bibliografía NLG. El número de sistemas y de dominios de aplicación es muy relevante, lo que hace que este ámbito se pueda considerar un campo de investigación maduro. Existe una metodología de desarrollo y validación generalmente aceptadas, aunque, dada la complejidad de muchos sistemas, en realidad suele ser el dominio de aplicación y sus necesidades específicas las que determinan como se deben realizar los diseños.

3. Descripciones Lingüísticas de Datos (LDD)

En el campo de la lógica borrosa y la computación flexible ha surgido un ámbito de trabajo orientado a la construcción de descripciones de datos mediante la utilización de términos lingüísticos [24]. Especialmente a partir de la propuesta del paradigma de la 'computación con palabras' [2] y su evolución hacia la 'computación con percepciones' surgió la denominada 'sumarización lingüística de datos' [24,25], que propone la utilización de sentencias cuantificadas borrosas sobre una o más variables como expresiones descriptivas de datos. Esta aproximación se ha aplicado posteriormente en diversos casos prácticos, pasando posteriormente a denominarse 'Descripción lingüística de Datos', ámbito que entiende los sumarios como una herramienta para la descripción de percepciones. La evidente limitación sintáctica de estos modelos ha dado lugar a que surjan propuestas que enriquezcan las posibilidades expresivas de este paradigma, lo que ha hecho que se consolide como un campo de investigación novedoso para la aplicación de modelos borrosos. La novedad de este ámbito se evidencia, además, en la inexistencia de metodologías de diseño y validación propias y bien

asentadas, así como en que, salvo recientes excepciones, las propuestas no suelen presentar aplicaciones prácticas, sino mayoritariamente ejemplos o casos de uso simples con diferentes niveles de complejidad.

Pese a lo anterior, sí resulta posible identificar algunos elementos fundamentales comunes a las soluciones LDD propuestas en la bibliografía. Así, en general se entiende el proceso de creación de LDD como la extracción de información utilizando términos lingüísticos imprecisos. En este sentido, esta tarea es similar a las que se utilizan en la etapa de 'Determinación de contenidos' de la metodología de desarrollo de los sistemas D2T, por lo que tiene un encuadre natural en dicha etapa, especialmente a la vista de la utilización generalizada de variables y valores lingüísticos y particiones borrosas. Otra característica común a gran parte de los sistemas LDD es la utilización de cuantificadores borrosos (tanto absolutos como relativos) para la generación de sentencias cuantificadas borrosas de tipo I (Q *son* A) y tipo II (Q *Bs son* A) como protoformas que describen los datos. Este tratamiento resulta más próximo a la etapa de lexicalización de los sistemas D2T, puesto que las descripciones obtenidas están dotadas de una cierta estructura sintáctica.

Una característica inherente al tratamiento borroso de la información es la de que cada a combinación de términos (descripciones candidatas) se le asocia un grado de cumplimiento que se calcula mediante los mecanismos de evaluación borrosa habituales (siguiendo los modelos de cuantificación borrosa que se utilicen en cada aplicación) lo que en si mismo constituye ya un criterio objetivo que mide la adecuación de la descripción a los datos. Adicionalmente existen otros criterios objetivos [26,27,28] como el grado de cobertura de los datos, la especificidad de los cuantificadores, la longitud de la descripción y otros que suelen utilizarse para determinar las descripciones mas adecuadas. De hecho sí existe un consenso sólido en que estos mecanismos son de utilidad para ordenar las descripciones candidatas de una manera objetiva.

Por último se han propuesto extensiones para dar mayor riqueza expresiva a los modelos, a través de referencias espacio-temporales, extensión a más de una variable o mediante la composición de múltiples sentencias, que han resultado de utilidad en algunos de los casos de uso descritos en la bibliografía, como:

- [28], donde se construyen LDD con proposiciones cuantificadas de tipo II sobre series de datos temporales, en base a criterios de brevedad, precisión y cobertura
- [29] orientada a la descripción de perfiles de patrones de evolución sobre datos de cotización del índice Nikkei
- [30] para la detección de patrones de variación en datos económicos

También se han propuesto ámbitos de aplicación real, como los descritos en:

- [31], para la generación de LDD relativas al consumo eléctrico doméstico
- [32] en el ámbito de la meteorología, para la predicción operativa individualizada por ayuntamientos en Galicia.

Otro aspecto donde D2T y LDD han seguido caminos separados es el proceso de evaluación y medición de la calidad y adecuación final de los sistemas. En general, los sistemas D2T llevan a cabo evaluaciones automáticas y (mayoritariamente) con expertos humanos, y que pueden ser cualitativas o cuantitativas. En el caso de las LDD, los criterios de evaluación han sido el único mecanismo para determinar la calidad de las descripciones. Pese a su utilidad, por si solos no tienen capacidad para proporcionar información sobre otros aspectos cruciales en la generación de narrativas. Por ejemplo, una LDD puede tener un valor de evaluación muy elevado pero, si su contenido es irrelevante para el usuario, el vocabulario es incorrecto o el texto es repetitivo o expresado con errores, la LDD no estará cumpliendo su objetivo. Tan importante como el *qué* resulta en este ámbito el *cómo* se transmite la información. En este sentido algunos autores del ámbito LDD han comenzado a utilizar técnicas y metodologías de evaluación propias de la D2T, como la evaluación basada en cuestionarios descrita en [33] o en [32], donde se tienen en cuenta las diferentes dimensiones que caracterizan una buena descripción de los datos.

LDD es un campo de investigación que se encuentra, por tanto, es sus etapas iniciales, con una sólida base formal y cuyo potencial es muy prometedor, pero está todavía por concretarse. Comienza a haber resultados relevantes, en aplicaciones complejas, pero la mayoría de las propuestas han sido formales o teóricas, tratando casos de uso muy simples. Por lo tanto, no puede decirse que las LDD tengan capacidad por sí solas para constituir narrativas que podamos denominar propiamente 'en lenguaje natural,' pero sí resultan adecuadas para producir expresiones básicas en lenguajes específicos. En realidad, las LDD tendrán éxito como paradigma en la medida en que logren evidenciar su valor como parte del proceso de un sistema D2T, como mecanismos capaces de gestionar la incertidumbre en la determinación del contenido de las narrativas o en su lexicalización.

4. Conclusiones

En este trabajo hemos revisado las dos principales aproximaciones que se utilizan para la generación de textos a partir de datos: sistemas 'Data-To-Text (D2T),' que, dentro de la Generación de Lenguaje Natural (NLG), tratan con el problema general de convertir datos en textos comprensibles y Descripción lingüística de datos (LDD), que trata con la abstracción de los datos en términos imprecisos. Ambas aproximaciones son en realidad complementarias, puesto que las abstracciones LDD sirven como métodos que pueden emplearse en la primera tarea que aborda la metodología de desarrollo de los sistemas D2T (determinación del contenido). En este contexto, la LDD aporta el valor adicional de incluir un grado de cumplimiento numérico de cada descripción, lo que permite ordenarlas y que puede gestionarse durante las siguientes etapas que constituyen el proceso D2T. Asimismo, las LDD con estructura sintáctica son elementos valiosos a ser considerados en la etapa de lexicalización. Más allá de las sentencias cuantificadas borrosas, otros modelos y metodologías de la Computación con

Palabras deberían dar lugar a vocabularios y sentencias de mayor complejidad que enriquezcan esta fase y puedan ser consideradas en las narrativas finales.

En este sentido, las propuestas ya existentes en el ámbito de la Computación con Palabras para el modelado de, entre otros, términos, relaciones, sentencias, así como para el razonamiento aproximado, son elementos que aportan un nuevo valor para la generación de descripciones sobre datos y en contextos con incertidumbre en el lenguaje, que pueden generalizar el tratamiento (en ocasiones muy básico) que se realiza actualmente en los sistemas D2T, así como también darle y nivel expresivo de mayor riqueza lingüística.

Por tanto, el desarrollo de aplicaciones que combinen ambas aproximaciones es una alternativa de indudable valor, que de hecho está comenzando a ser considerada. Existe un cierto consenso en que la viabilidad de las LDD como campo de investigación exitoso que pueda aplicarse a problemas del mundo real está ligada a esta integración como parte de sistemas D2T. Más aún, en dominios donde las fuentes de datos sean esencialmente numéricas (series temporales, por ejemplo) los métodos y tecnologías LDD resultan más adecuados por su capacidad para gestionar términos con incertidumbre en las expresiones lingüísticas.

Agradecimientos Esta investigación ha sido financiada por el Ministerio de Economía y Competitividad (proyecto TIN2014-56633-C3-1-R, cofinanciado por el Programa FEDER) y por la Xunta de Galicia (proyectos EM2014/012 y CN2012/151, cofinanciados por el Programa FEDER). A. Ramos-Soto está financiado por el Ministerio de Economía y Competitividad mediante el programa de contratos predoctorales FPI.

Referencias

1. Bateman, J.A.: Natural language generation: an introduction and open-ended review of the state of the art, http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/nlg/atg01/node1.html (2001)
2. Zadeh, L.A.: Fuzzy logic = computing with words. Fuzzy Systems, IEEE Transactions on **4**(2) (1996) 103–111
3. Zadeh, L.A.: From computing with numbers to computing with words : From manipulation of measurements to manipulation of perceptions. In: Intelligent Systems and Soft Computing: Prospects, Tools and Applications, Springer-Verlag (2000) 3–40
4. Kacprzyk, J.: Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation. IEEE Trans. Fuzzy Systems (2010) 451–472
5. Bateman, J.A., Zock, M.: Nlg systems wiki, http://www.nlg-wiki.org/systems/
6. Ramos-Soto, A., Bugarín, A., Barro, S.: On the role of linguistic descriptions of data in the building of natural language generation systems. Fuzzy Sets and Systems (2015)
7. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press (2000)

8. Reiter, E., Dale, R.: Building applied natural language generation systems. Journal of Natural-Language Engineering (3) (1997) 57–87
9. Goldberg, E., Driedger, N., Kittredge, R.: Using natural-language processing to produce weather forecasts. IEEE Expert **9**(2) (1994) 45–53
10. Coch, J., Dycker, E.D., García-Moya, J.A., Gmoser, H., Stranart, J.F., Tardieu, J.: Multimeteo: adaptable software for interactive production of multilingual weather forecasts. In: Proceedings of the 4th European Conference on Applications of Meteorology (ECAM 99), Norrköping, Sweden (1999)
11. Sripada, S., Reiter, E., Davy, I.: Sumtimemousam: Configurable marine weather forecast generator. Expert Update **6**(3) (2003) 4–10
12. Sripada, S.G., Burnett, N., Turner, R., Mastin, J., Evans, D.: A case study: Nlg meeting weather industry demand for quality and quantity of textual weather forecasts. In: INLG-2014 Proceedings. (June 2014)
13. Busemann, S., Horacek, H.: Generating air-quality reports from environmental data. In Busemann, S., Becker, T., Finkler, W., eds.: DFKI Workshop on Natural Language Generation, DFKI Document D-97-06. (1997)
14. Wanner, L., Bohnet, B., Bouayad-Agha, N., Lareau, F., Nicklass, D.: Marquis: Generation of user-tailored multilingual air quality bulletins. Applied Artificial Intelligence **24**(10) (2010) 914–952
15. Hüske-Kraus, D.: Suregen 2: A shell system for the generation of clinical documents. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003). (2003) 215–218 (Research Notes and Demos).
16. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artif. Intell. **173**(7-8) (May 2009) 789–816
17. Cogentex, I.: Project reporter website, http://www.cogentex.com/products/reporter/
18. Sheremetyeva, S., Nirenburg, S., Nirenburg, I.: Generating patent claims from interactive input. In: Proceedings of the 8th. International Workshop on Natural Language Generation (INLG'96), Herstmonceux, England (June 1996) 61–70
19. Yu, J., Hunter, J., Reiter, E., Sripada, S.: An approach to generating summaries of time series data in the gas turbine domain. In: Proceedings of IEEE International Conference on Info-tech & Info-net (ICII2001), Beijing (2001) 44–51
20. Belz, A., Reiter, E.: Comparing automatic and human evaluation of nlg systems. In: In Proc. EACL'06. (2006) 313–320
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 311–318
22. Sambaraju, R., Reiter, E., Logie, R., McKinlay, A., McVittie, C., Gatt, A., Sykes, C.: What is in a text and what does it do: Qualitative evaluations of an nlg system – the bt-nurse – using content analysis and discourse analysis. In: Proceedings of the 13th European Workshop on Natural Language Generation. (2011) 22 – 31
23. Reiter, E.: Task-based evaluation of nlg systems: Control vs real-world context. In: Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop. UCNLG+EVAL '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 28–32
24. Yager, R.R.: A new approach to the summarization of data. Information Sciences **28**(1) (1982) 69 – 86

25. Yager, R.R., Ford, K.M., Cañas, A.J.: An approach to the linguistic summarization of data. In Bouchon-Meunier, B., Yager, R.R., Zadeh, L.A., eds.: Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 90, Paris, France, July 2-6, 1990, Proceedings. Volume 521 of Lecture Notes in Computer Science., Springer (1990) 456–468
26. Díaz-Hermida, F., Ramos-Soto, A., Bugarín, A.: On the role of fuzzy quantified statements in linguistic summarization. In: Proceedings of 11th International Conference on. Intelligent Systems Design and Applications (ISDA). (2011) 166–171
27. Castillo-Ortega, R., Marín, N., Sánchez, D., Tettamanzi, A.: Quality assessment in linguistic summaries of data. In Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R., eds.: Advances in Computational Intelligence. Volume 298 of Communications in Computer and Information Science. Springer Berlin Heidelberg (2012) 285–294
28. Castillo-Ortega, R., Marín, N., Sánchez, D.: A fuzzy approach to the linguistic summarization of time series. Multiple-Valued Logic and Soft Computing (2011) 157–182
29. Kobayashi, I., Okumura, N.: Verbalizing time-series data: With an example of stock price trends. In: Proceedings IFSA/EUSFLAT Conf. (2009) 234–239
30. Kacprzyk, J., Wilbik, A.: Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In: Proceedings IFSA/EUSFLAT Conf. 2009. (2009) 1321–1326
31. van der Heide, A., Trivino, G.: Automatic generated linguistic summaries of energy consumption data. In: Proceedings of 9th ISDA Conference. (2009) 553–559
32. Ramos-Soto, A., Bugarín, A., Barro, S., Taboada, J.: Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. Fuzzy Systems, IEEE Transactions on **23**(1) (2015) 44 – 57
33. Eciolaza, L., Pereira-Fariña, M., Trivino, G.: Automatic linguistic reporting in driving simulation environments. Applied Soft Computing **13**(9) (2013) 3956 – 3967