ICA2TEXT: Un sistema para la descripción automática en lenguaje natural de series temporales de calidad del aire

ICA2TEXT: A system for the automatic natural language description of air quality time series

Andrea Cascallar-Fuentes[1], Javier Gallego-Fernández[1], Alejandro Ramos-Soto[1], Anthony Saunders-Estévez[2] and Alberto Bugarín-Diz[1]

[1]*Grupo de Sistemas Intelixentes, Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez s/n, Campus Vida 15782, Santiago de Compostela, España*

[2]*Rede de Calidade do Aire de Galicia, MeteoGalicia, Xunta de Galicia, Calle Roma 6 15707 Fontiñas, Santiago de Compostela, España*

Resumen

En este proyecto describimos ICA2TEXT, un sistema data-to-text para generar automáticamente descripciones textuales sobre series temporales de calidad del aire proporcionadas por MeteoGalicia. Los resultados de la evaluación por parte de dos expertos meteorólogos fueron muy satisfactorios, lo que confirma que las descripciones textuales propuestas se ajustan a este tipo de datos y servicios tanto en contenido como en diseño. Actualmente, este sistema se encuentra en una fase final de pruebas y será desplegado como servicio público de la web de MeteoGalicia [1].

English translation. In this project we describe ICA2TEXT, a data-to-text system to automatically generate textual descriptions about air quality time series provided by MeteoGalicia. Assessment results by two experts meteorologists were very satisfactory, which confirm that the proposed textual descriptions fit this type of data and service both in content and layout. This system is currently in a final testing phase and will be deployed as a public service on the MeteoGalicia website [1].

Keywords

términos lingüísticos borrosos, sistemas data-to-text, generación de lenguaje natural.

1. Introducción

Profundizar en la información realmente relevante que hay detrás de los datos plantea la necesidad de emplear técnicas que se adapten a las necesidades específicas de cada dominio y que puedan escalar a medida que se acumulan los datos.

La Generación de Lenguaje Natural (NLG) es un campo centrado en la generación de texto a partir de varias fuentes de datos. Dentro del NLG, los sistemas data-to-text (D2T) [2] generan automáticamente textos a partir de grandes conjuntos de datos numéricos o simbólicos, proporcionando información comprensible. Normalmente, el diseño de los sistemas D2T incluye *i)* una etapa de análisis de datos donde se extrae la información relevante y *ii)* una etapa de generación donde se transmite la información en lenguaje natural. Relacionado con esto, desde el campo de la lógica borrosa se ha propuesto varios enfoques para generar descripciones lingüísticas de los datos (LDD) o resúmenes lingüísticos utilizando términos lingüísticos.

En este trabajo describimos ICA2TEXT, un sistema data-to-text basado en la lógica borrosa y la generación de lenguaje natural para describir automáticamente series temporales sobre el índice de calidad del aire (ICA), que es un indicador ampliamente utilizado en todo el mundo de la calidad del aire.

2. Contexto del problema

La presencia de contaminantes en el aire y, por tanto, el deterioro de la calidad del aire puede tener efectos nocivos para la salud de las personas. Hemos trabajado con datos describen el Índice de Calidad del Aire (ICA) en la red de 50 estaciones meteorológicas que envían datos actualizados cada hora en tiempo real en Galicia proporcionados por MeteoGalicia [1]. Para determinar la calidad del aire, este servicio mide cinco contaminantes diferentes: SO_2, NO_2, $PM25$, $PM10$ and O_3.

Basándose en los criterios de la Agencia Europea de Medio Ambiente [3], esta variable tiene seis etiquetas con una percepción positiva, neutra o negativa (Tabla 1).

Debido a la importancia de esta información, los meteorólogos de MeteoGalicia pretenden ofrecerla a los ciudadanos de forma comprensible, hasta ahora en formato gráfico. Por ello, surge la necesidad de dotar a esta in-

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

andrea.cascallar.fuentes@usc.es (A. Cascallar-Fuentes); javier.gallego.fernandez@rai.usc.es (J. Gallego-Fernández); alejandro.ramos@inverbisanalytics.com (A. Ramos-Soto); calidadedoaire.cma@xunta.gal (A. Saunders-Estévez); alberto.bugarin.diz@usc.es (A. Bugarín-Diz)

0000-0003-1857-5796 (A. Cascallar-Fuentes); 0000-0001-6136-8413 (A. Ramos-Soto); 0000-0003-3574-3843 (A. Bugarín-Diz)

Tabla 1
Etiquetas del índice de calidad del aire con su percepción e índice numérico.

Percepción	Positiva		Neutra		Negativa	
Etiqueta	Muy bueno	Bueno	Regular	Malo	Muy malo	Pésimo
Índice	0	1	2	3	4	5

formación gráfica de una descripción textual que facilite su comprensión. En este contexto, hemos desarrollado el sistema ICA2TEXT en colaboración con los expertos de MeteoGalicia para describir lingüísticamente las series temporales de calidad del aire. El diseño de este sistema ha sido realizado de modo que atiende a las necesidades de este ámbito en cuanto a la flexibilidad de la riqueza lingüística requerida, abordando el manejo de la imprecisión en la descripción de series temporales. En los siguientes apartados se muestra en detalle el diseño del sistema siguiendo los requerimientos de los expertos.

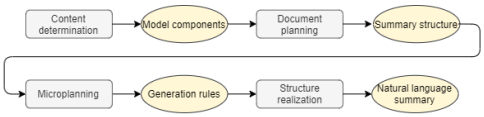

Figura 1: Representación de la arquitectura de nuestra propuesta. Los rectángulos representan las etapas, mientras que las elipses representan los resultados.

3. Descripciones lingüísticas de las series temporales del ICA

Este sistema se compone de las siguientes etapas (Figura 1), que componen la arquitectura data-to-text propuesta para describir las series temporales.

3.1. Determinación de contenido

Esta fase se compone de dos sub-etapas: *i)* Análisis de los datos, en el que se identifican los patrones y las tendencias, e *ii)* interpretación de los datos, en la que se identifican los mensajes que representan los patrones y la relación entre ellos.

Hemos diseñado un modelo temporal borroso para abordar el problema de manejar la imprecisión de la información temporal al resumir las series temporales. Este modelo temporal se ha diseñado para agrupar los datos, si es posible, en la referencia temporal más general. Nuestro objetivo es que el discurso sea legible y comprensible, aunque se pierda algo de precisión o exactitud en las descripciones.

3.2. Planificación del documento

Una vez identificados los mensajes y sus relaciones, en esta fase se generan todos los mensajes que se pueden incluir en la descripción final y se da una estructura a la descripción lingüística. La estructura de la descripción lingüística es la siguiente: *i* resumen general, *ii* intensificación (si procede) y *iii* excepción (si procede). Además, el resumen general incluye una descripción general y la descripción de la tendencia si procede, mientras que las secciones de intensificación y excepción contienen valores excepcionales ordenados de forma ascendente por valor o fecha. Realizamos las descripciones lingüísticas en los idiomas español y gallego utilizando SimpleNLG-ES [4] y SimpleNLG-GL [5].

3.3. Microplanificación

A partir de los mensajes generados previamente y de la estructura definida, en esta fase se seleccionan los casos a destacar y, por tanto, los mensajes que se van a mostrar. Las reglas de microplanificación se basan en las máximas griceanas [6].

En cuanto al resumen general se define que *i)* al describir un caso negativo se debe incluir el contaminante causante y *ii)* la tendencia sólo se incluye si las etiquetas de inicio y fin son diferentes.

En cuanto a la intensificación y a la excepción se define que *i)* el contaminante causante de un ICA negativo se omite si ha sido indicado en el resumen general, *ii)* se debe seleccionar la referencia temporal más general posible con un grado de verdad mayor o igual a 0.9 y *iii)* los periodos de tiempo con el mismo valor se agrupan en la descripción.

3.4. Realización de la estructura

Una vez que hemos definido la estructura y los mensajes que compondrán la descripción lingüística, se genera automáticamente asegurándonos de que sea correcta ortográfica, morfológica y sintácticamente. En este escenario, tanto en la intensificación como en la excepción, si el número de casos destacados es superior a 2, se dispondrán como una lista. Sin embargo, cuando el número de elementos sea igual o inferior a 2 se incluirán ambos como texto plano.

3.5. Definición de los componentes

En esta sección, presentamos el diseño de los componentes necesarios para generar la descripción lingüística de la serie de índices de calidad del aire.

3.5.1. Cálculo de las etiquetas

En primer lugar, calculamos la etiqueta del índice general de calidad del aire que mejor representa la serie temporal global para incluirla en la descripción general. Esta etiqueta se obtuvo como una media ponderada en la que el valor más reciente es el más relevante para describir la situación general a través de la referencia temporal "En las últimas horas". Además, en descripción de la tendencia, su valor también se calculaba con una media ponderada.

3.5.2. Referencias temporales

En el libro de estilo de MeteoGalicia, se define la franja horaria para las diferentes partes del día {mañana, tarde, noche} en verano e invierno.

Aunque los rangos que definen estos momentos del día se declaran de forma estática (al igual que la definición de un día completo desde las 00:00:00 hasta las 23:59:59), su uso al hablar está condicionado por la imprecisión del lenguaje. De modo que hemos definido de forma difusa las siguientes referencias temporales:

- Día completo: en lugar de una definición estricta desde las 00:00:00 hasta las 23:59:59, agrupamos como día también las dos horas anteriores y posteriores con un peso en el rango [0, 1].
- Mañana, tarde, noche: como se ha mencionado anteriormente, estas referencias temporales están definidas en el libro de estilo de MeteoGalicia. Utilizando esa definición como base, las hemos definido como un conjunto borroso trapezoidal en el que las dos horas anteriores y posteriores a los límites se consideran con un peso en el rango [0, 1].
- Primeras, centrales y últimas horas de la {mañana, tarde y noche}: hemos definido estas tres referencias temporales para describir situaciones más específicas. Estas etiquetas también se definen como conjuntos borrosos trapezoidales.

4. Validación por expertos

Hemos pedido a dos meteorólogos expertos de la Red de Calidad del Aire de MeteoGalicia [1] que evaluaran la calidad de las descripciones lingüísticas generadas por ICA2TEXT en este dominio y su adecuación rellenando el cuestionario compuesto por 30 situaciones meteorológicas diferentes utilizando una escala de 5 puntos donde 1 significa "el experto está absolutamente en desacuerdo" y 5 "el experto está absolutamente de acuerdo". Ninguno de estos dos expertos había participado en la definición de ninguna parte del modelo.

Es cuestionario está formado por cinco preguntas, agrupadas en dos categorías: contenido de la descripción lingüística (Q1, Q2) y diseño (Q3, Q4, Q5). Cada caso del cuestionario está formado por una representación gráfica de la serie temporal y la descripción textual generada que describía el caso, pidiéndoles que evaluaran la idoneidad de las descripciones para describir las distintas situaciones. La figura 2 muestra un ejemplo extraído del cuestionario.

Tabla 2
Preguntas del cuestionario de validación de expertos del índice de calidad del aire.

Código	Pregunta
Q1	La descripción lingüística representa correctamente los datos representados en la figura
Q2	La descripción concuerda con la forma en que describirías los datos
Q3	El vocabulario se usa correctamente
Q4	La organización de la descripción lingüística facilitar su comprensión
Q5	La ortografía, la puntuación y la estructura son correctas

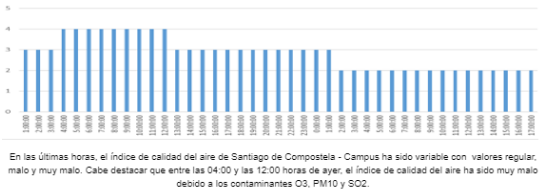

Figura 2: Ejemplo del cuestionario de evolución del ICA diseñado para la validación de expertos.

Tabla 3
Resultado de la evaluación realizada por expertos.

	Media	Desv. Típica	Moda	Mediana	IQR
Q1	4.58	0.87	5	5	1
Q2	4.15	1.01	5	4	1
Q3	4.75	0.70	5	5	0
Q4	4.92	0.28	5	5	0
Q5	4.97	0.18	5	5	0
Contenido	4.37	0.96	5	5	1
Estructura	4.88	0.46	5	5	0
General	4.67	0.75	5	5	0

En la Tabla 3 presentamos un resumen de las puntuaciones de los expertos para cada una de las preguntas de forma individual y agrupada por dimensión. En general, los resultados muestran que los expertos están de acuerdo con las descripciones lingüísticas, ya que la media de las puntuaciones es de 4,67 y la moda muestra que el mayor valor utilizado es 5, es decir, la puntuación máxima. Por lo tanto, podemos concluir que estas descripciones lingüísticas generadas son muy adecuadas tanto en contenido como en forma para describir series temporales de índices de calidad del aire.

5. Discusión y conclusiones

En este trabajo hemos descrito el desarrollo de ICA2TEXT, un sistema que genera descripciones lingüísticas de datos de calidad del aire en castellano y gallego en colaboración con expertos de MeteoGalicia. Nuestro objetivo era cubrir las necesidades detectadas de acompañar la información gráfica que ofrecen en su web con descripciones textuales que faciliten su comprensión por parte de los usuarios.

Las series temporales para cada estación nunca supera los 150 registros. Nuestra aproximación consume una media de 10s para generar las dos descripciones textuales (una por idioma) para las 50 estaciones de MeteoGalicia. Este tamaño es lo usual por lo que nuestra aproximación puede ser utilizada con datos de cualquier agencia meteorológica realizando las adaptaciones pertinentes.

ICA2TEXT permite incluir un nuevo idioma, incluyendo los elementos necesarios en los archivos de configuración. Para los idiomas para los que ya existe una versión de SimpleNLG se podría adaptar fácilmente teniendo en cuenta las características de cada idioma. En caso de que no exista, habría que crear plantillas o un realizador lingüístico para este idioma.

Con respecto a su reutilización con otro tipo de datos, a la hora de describir series temporales se utiliza un tipo de relato muy habitual, donde se describe una valoración general de una situación incluyendo matices de intensificación y excepción. En el modelo que hemos definido hemos seguido esta estructura, de modo que, para reutilizar ICA2TEXT con otros tipos de datos, debería adaptarse la fase de preprocesado de los datos y las tareas realizadas dentro de la fase de determinación de contenido. Por otro lado, en caso de que los requisitos del lenguaje sean muy diferentes, habría que adaptar todas la fases del diseño en gran medida.

Los resultados de la validación realizada por expertos en la materia han sido muy satisfactorios. Como consecuencia, actualmente está siendo sometido a una fase final de pruebas y se desplegará como servicio público en la web oficial de MeteoGalicia.

Como trabajo actual y futuro, estamos aplicando nuestro modelo al diseño de nuevos sistemas D2T en otros ámbitos, como la notificación automática de series temporales en el ámbito de la sanidad electrónica.

Agradecimientos

Esta investigación ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades (subvenciones TIN2017-84796-C2-1-R, PID2020-112623GB-I00, y PDC2021-121072-C21) y la Consellería de Educación, Universidade e Formación Profesional (subvenciones ED431C2018/29 y ED431G2019/04). Todas las subvenciones han sido cofinanciadas por el Fondo Europeo de Desarrollo Regional (programa FEDER).

Referencias

[1] MeteoGalicia, MeteoGalicia website, 2021. URL: www.meteogalicia.gal, [Accessed February 2021].

[2] E. Reiter, An architecture for data-to-text systems, in: Proceedings of the Eleventh European Workshop on Natural Language Generation, Association for Computational Linguistics, 2007, pp. 97–104. URL: https://doi.org/10.3115%2F1610163.1610180. doi:10.3115/1610163.1610180.

[3] European Environment Agency, European Air Quality Index website, 2021. URL: www.eea.europa.eu, [Accessed February 2021].

[4] A. Ramos-Soto, J. J. Gallardo, A. Bugarín, Adapting SimpleNLG to Spanish, in: Proceedings of the 10th International Conference on Natural Language Generation, INLG, Association for Computational Linguistics, 2017, pp. 144–148. URL: https://doi.org/10.18653/v1/w17-3521. doi:10.18653/v1/w17-3521.

[5] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, Adapting SimpleNLG to Galician language, in: Proceedings of the 11th International Conference on Natural Language Generation, Association for Computational Linguistics, 2018, pp. 67–72. URL: https://doi.org/10.18653/v1/w18-6507. doi:10.18653/v1/w18-6507.

[6] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.