Modelado borroso de referencias geográficas textuales sobre datos de expertos

A. Ramos-Soto*†, Jose M. Alonso*, Ehud Reiter†, Kees van Deemter†‡, Albert Gatt§

* Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela
alejandro.ramos@usc.es
josemaria.alonso.moral@usc.es

† Department of Computing Science, University of Aberdeen
alejandro.soto@abdn.ac.uk
e.reiter@abdn.ac.uk
k.vdeemter@abdn.ac.uk

‡ Department of Information and Computing Sciences, Utrecht University
k.vandeemter@uu.nl

§ Institute of Linguistics and Language Technology, University of Malta
albert.gatt@um.edu.mt

***Resumen*—Describimos una metodología para la construcción de modelos borrosos de expresiones geográficas sobre interpretaciones individuales de dichas expresiones por parte de expertos. Esta metodología aborda una tarea de modelado de lenguaje encuadrado dentro del desarrollo de un sistema *data-to-text* que generará descripciones textuales sobre mapas con información meteorológica en tiempo real. Para ello, recogimos datos proporcionados por meteorólogos en una encuesta y, sobre los mismos, creamos modelos borrosos consistentes que agregan las diferencias de interpretación de los expertos. Estos modelos permitirán generar expresiones de referencia geográfica sobre eventos meteorológicos georreferenciados, que formarán parte de los textos generados por el sistema *data-to-text* que desarrollaremos en un trabajo futuro.**

***Términos clave*—vaguedad, generación de lenguaje natural, conjuntos borrosos, modelado de lenguaje, sistemas de información geográfica, descripciones lingüísticas de datos**

I. Introducción

La ingente cantidad de datos que se producen en todo tipo de ámbitos ha favorecido la aparición de varias disciplinas centradas en investigar cómo proporcionar a las personas la información relevante que permanece latente en dichos datos. Una de estas disciplinas es la generación de lenguaje natural (NLG), que estudia el problema de cómo generar textos a partir de datos, que puedan resultar útiles a lectores humanos [1], [2]. Dentro de este campo, los sistemas que generan textos a partir de datos no lingüísticos se conocen como *data-to-text* (D2T). En los últimos tiempos este tipo de sistemas gozan de cierta popularidad, gracias a su amplio uso comercial [2], [3] en una gran variedad de dominios.

Gran parte de los sistemas D2T proporcionan textos o informes que describen series de datos temporales, y ciertamente se pueden encontrar muchos ejemplos en el estado del arte, e.g., en predicción meteorológica [4]–[7], salud [8]–[10], o industria [11], entre muchos otros [1], [12]. Así mismo, también existen sistemas D2T que tratan datos caracterizados geográficamente, aunque su número es mucho más reducido respecto a los alimentados por series temporales de datos [13].

Una de las tareas esenciales en la concepción de un sistema D2T es el modelado del lenguaje, i.e., definir la semántica de los términos y expresiones a usar para describir los datos [14], [15]. Existen distintas formas de resolver esta tarea, tales como usar algoritmos heurísticos o de aprendizaje máquina sobre un conjunto de *corpus* paralelo de texto y datos [16] para crear modelos de las expresiones de interés [17], conseguir que los expertos proporcionen dichos modelos, o recolectar datos de escritores o lectores que puedan ser usados para la aplicación de algoritmos de mapeado.

Por ejemplo, los sistemas D2T que generan textos a partir de series temporales de datos incluyen generalmente expresiones temporales para referirse a eventos o patrones relevantes encontrados en los datos. Realizar la tarea de modelado de lenguaje sobre expresiones temporales en este tipo de sistemas permite asegurarnos de que los textos generados incluyen términos cuyo significado se alinea con la interpretación de los expertos o las expectativas de los lectores, [17], [18]. Del mismo modo, disponer de un buen modelo geográfico en sistemas que generan textos sobre datos geográficos es esencial para generar expresiones que se refieran a localizaciones concretas y a regiones de interés.

Al mismo tiempo, es bastante frecuente que las expresiones temporales y geográficas que deben ser incluidas en los textos generados por sistemas D2T sean vagas, como “por la tarde” [4] o “áreas del suroeste” [19]. En situaciones donde la vaguedad aparece (y con ella, casos fronterizos y conceptos graduales), se ha propuesto el uso de conjuntos borrosos para modelar expresiones lingüísticas en sistemas D2T [2], [20], [21]. Sin embargo, no existen sistemas de este tipo en el estado del arte que hagan uso de dichas técnicas, con la excepción de GALiWeather [5], que hace un uso básico de conjuntos borrosos para modelar expresiones temporales y cuantificadores.

En este contexto, este trabajo describe la metodología que hemos seguido para llevar a cabo una tarea de modelado de lenguaje de expresiones geográficas vagas. Esta tarea forma parte de un proyecto más amplio cuyo objetivo es la creación de un sistema D2T que genere descripciones textuales de mapas con información meteorológica en tiempo real. Nuestra aproximación se compone de una tarea de recogida de datos, proporcionados por expertos, y de un algoritmo heurístico que agrega dichos datos para crear modelos borrosos.

II. Trabajo relacionado

El campo de NLG es amplio y existen gran cantidad de tipos de sistemas existentes con distintos propósitos como la gen-

eración de informes a partir de datos, creación de resúmenes a partir de distintas fuentes textuales, o generación de diálogo, narrativas e incluso poesía [1]. En nuestro caso, nos centramos en un problema muy específico, aquellos sistemas D2T cuyos datos de entrada están caracterizados geográficamente y cuyos textos a generar incluyen expresiones geográficas que se refieren a la ocurrencia de ciertos eventos registrados en los datos (e.g., "lluvia en el norte de España", "inundaciones en la Costa del Sol" o "vientos fuertes en el noroeste de Escocia").

Aunque este tipo de referencias fueron introducidas hace décadas en los textos generados por el sistema FoG [6], RoadSafe es quizás el ejemplo más representativo y reciente de este tipo de sistemas [19], [22]. RoadSafe usaba datos de predicción meteorológica para generar predicciones textuales orientadas al mantenimiento de carreteras. Estos informes incluían expresiones temporales y geográficas para ayudar a identificar dónde y cuándo ciertos fenómenos relevantes tendrían lugar, con el objetivo de ayudar a los equipos de mantenimiento a mantener las condiciones de las vías afectadas en buen estado (ver ejemplos de dichas expresiones en Fig. 1). Así pues, desarrollar RoadSafe supuso también un estudio profundo de cómo generar buenas expresiones geográficas que se refiriesen de forma adecuada a la geografía subyacente a la información relevante extraída de los datos de entrada.

Concretamente, la aproximación de RoadSafe para modelar la generación de expresiones de referencia geográficas se basaba en técnicas estándar del campo de sistemas de información geográfica (GIS), en el que particionan la geografía subyacente de los eventos utilizando distintos esquemas, o *marcos de referencia* espaciales [23], que a su vez se componen de particiones no solapables (conocidas como *descriptores*). Por ejemplo, el marco de referencia *Dirección* se compone de los descriptores "nordeste", "suroeste", etc., y el marco "Proximidad Costera" se compone de los descriptores "costa" e "interior".

Una vez que los límites numéricos de cada descriptor se definen para cada marco usando coordenadas de latitud-longitud, cada punto de datos puede ser caracterizado por un conjunto de descriptores (ej. "suroeste" y "costa") y el generador de expresiones de referencia geográficas se encarga de seleccionar el mejor conjunto de descriptores que describan el área formada por el subconjunto de puntos que representan el evento.

En trabajos más recientes, el modelo geográfico utilizado en RoadSafe fue ampliado mediante la inclusión de referencias espaciales de nombres propios, que según un estudio de varios conjuntos de *corpus* en distintos dominios, son las más predominantes [13]. Además, se plantea el desarrollo de un algoritmo de generación de expresiones de referencia geográficas que integre el marco de referencia de nombres propios con los ya existentes en RoadSafe. Por otro lado, en [24] se tratan las diferencias entre referencias absolutas y relativas y se proporciona un modelo basado en la mereología, en la que los descriptores de nombre propio se combinan con descriptores de otros marcos de referencia.

Las referencias descritas hasta el momento proporcionan buenas aproximaciones que permiten generar expresiones de referencia geográfica apropiadas sobre un conjunto de marcos de referencia. Sin embargo, en todos los casos los modelos descritos fueron desarrollados basados en un particionado nítido de la geografía a tratar. Tomar límites exactos entre descriptores geográficos puede considerarse una asunción poco intuitiva, especialmente si tenemos en cuenta cómo las personas entendemos y usamos incluso las referencias geográficas más simples como "norte" u "oeste", en los que realmente las fronteras no suelen estar bien definidas, sino que son vagas. Por ejemplo, usando cualquiera de las aproximaciones anteriores, si una aldea se encuentra situada en la frontera entre dos descriptores, dependiendo de la granularidad espacial de nuestros datos es posible asignar descriptores opuestos (como "norte" y "sur") a dos puntos dentro de la misma localidad.

Por ello, las limitaciones que en este sentido presentan los modelos descritos anteriormente suponen un fuerte incentivo a la hora de buscar otro tipo de aproximaciones que permitan modelar la imprecisión o la incerteza en el uso de referencias geográficas vagas. De hecho, el problema del tratamiento de la vaguedad en referencias geográficas no se limita únicamente a sistemas D2T, sino que existe una discusión más amplia dentro del ámbito de GIS, desde hace ya varias décadas [25]. En este sentido, la teoría de conjuntos borrosos ha sido aplicada en numerosos casos para tratar la vaguedad en conceptos geográficos y relaciones espaciales [26], [27].

Del mismo modo, como se ha descrito en la Sec. I, de una manera más general también se ha propuesto el uso de conjuntos borrosos en sistemas D2T para modelar términos vagos [2], [20], [21]. A día de hoy, el único sistema D2T desplegado en un entorno real que hace un uso básico de este tipo de técnicas es GALiWeather [5], si bien existe un número importante de casos de uso de aplicación de conjuntos borrosos para extracción de información lingüística (descripciones lingüísticas de datos), que en ocasiones se acompaña de generación textual basada en plantillas [2], [12].

Road surface temperatures will fall below zero **during the late evening and tonight** except in **areas below 100M**.

SW 10-25 gusts **this afternoon in southwestern areas**, veering WSW and increasing 15-35 after midnight, gusts 55-60 **during the evening and tonight** except **in areas above 500M**, increasing 20-45 then veering W **by early morning**, gusts 70-75 **tomorrow morning in most southern and central places**.

Wintry precipitation will affect most routes at first, falling as snow flurries **in some places above 300M** at first. Snow spreading **throughout the forecast period** to **all areas** and persisting **in some places above 300M** until end of period.

Fig. 1. Ejemplos de textos generados por RoadSafe [19]

III. Motivación

Partiendo de las limitaciones que presentan los desarrollos previos de sistemas D2T que generan expresiones de referencia geográfica, la principal motivación de este trabajo es mejorar el modelado de conceptos geográficos vagos para fines de generación de lenguaje natural. Concretamente, nuestro objetivo es establecer una metodología de creación de modelos de

referencias geográficas vagas y de algoritmos de generación de expresiones de referencia sobre dichos modelos, basados en el uso de técnicas de conjuntos borrosos.

Puesto que D2T es un campo eminentemente aplicado, en el que los avances a nivel de investigación vienen dados por una necesidad real, en nuestro caso la metodología que proponemos se enmarca dentro del desarrollo de un sistema D2T para la generación de descripciones del estado meteorológico en tiempo real, sobre datos proporcionados por la Agencia de Meteorología de la Xunta de Galicia, MeteoGalicia [28]. Dichas descripciones incluirán expresiones de referencia geográfica que permitirán identificar fenómenos meteorológicos relevantes en el mapa, tales como temperatura, viento y estado del cielo.

En el marco del desarrollo del sistema D2T propuesto, este trabajo describe la tarea de modelado de lenguaje de las expresiones geográficas de interés a incluir en los textos generados por el sistema. Por otro lado, la metodología aquí descrita se basa en las ideas propuestas en [29] y [30], lo que permitirá consolidar nuevas formas de aplicar la teoría de conjuntos borrosos en sistemas D2T de forma práctica.

IV. Recogida de datos de expertos

Aunque es corriente realizar el modelado de lenguaje partiendo de un conjunto paralelo de textos y datos, con el fin de analizar el significado de las palabras y expresiones a modelar, en nuestro caso este tipo de recurso no se encontraba disponible. Por un lado, no disponíamos de acceso a un conjunto de datos extenso (ej. predicciones textuales y datos de predicción) en los que realizar tal análisis. Por otro, nuestro plan es desarrollar un nuevo sistema D2T para proporcionar descripciones textuales de datos meteorológicos en tiempo real. Así pues, la aproximación que tomamos consistió en interactuar de forma directa con los expertos.

A. La encuesta

Dado que nuestro propósito es el modelado de expresiones geográficas, pedimos directamente al director del departamento de predicción de *MeteoGalicia* una lista con las expresiones geográficas más usadas por los meteorólogos a la hora de escribir predicciones textuales. Tomando esta lista como base, preparamos una encuesta web que fue distribuida entre los expertos de la agencia meteorológica. En dicha encuesta, se pidió a los participantes que, sobre un mapa de la región de Galicia (mostrada bajo una proyección *Mercator*), dibujasen un polígono que representase una referencia geográfica dada (ver la Figura 2).

Fig. 2. Captura de pantalla de la encuesta realizada por los meteorólogos.

Marco de referencia		Descriptores
Dirección cardinal		Norte de Galicia, Sur de Galicia, Oeste de Galicia, Este de Galicia, Tercio norte, Extremo norte, Noroeste de Galicia, Noreste de Galicia, Suroeste de Galicia, Sureste de Galicia
Costa / Interior		Interior de Galicia
Nombre propio		Rías Baixas, Comarcas atlánticas
Mixto	Nombre y dirección	Oeste de A Coruña, Oeste de Ourense, Sur de Ourense, Sur de Lugo
	Nombre y costa	Litoral Atlántico, Litoral Cantábrico, Litoral norte, Interior de Coruña, Interior de Pontevedra
	Nombre y elevación	Áreas de montaña de Lugo, Áreas de montaña de Ourense

TABLA I
Lista de referencias geográficas de la encuesta.

En la encuesta se proporcionó a los participantes una lista de 24 descriptores, que aparecían en orden aleatorio. En esta lista, 20 de los 24 descriptores componen la lista original proporcionada por el director de predicción operativa, e incluyen direcciones cardinales, nombres propios y otro tipo de referencias como áreas montañosas, partes de provincias, etc. (ver la Tabla I para una taxonomía completa de los descriptores). Los restantes cuatro fueron añadidos para estudiar la combinación de direcciones cardinales mediante intersección (ej. explorar formas de combinar "norte" y "oeste" para obtener un modelo similar a "noroeste"), pero para nuestro propósito actual nos interesan sólo aquellos utilizados por los meteorólogos a la hora de escribir predicciones.

B. Resultados

La encuesta fue contestada por ocho expertos, obteniéndose 192 polígonos en total (160 sin considerar las intersecciones cardinales). A un nivel general, habíamos hipotetizado que los expertos serían bastante consistentes, dada su experiencia profesional. Así mismo, esperábamos también cierta variación entre las distintas respuestas.

Observamos que dichas hipótesis se han cumplido con claridad; los polígonos dibujados por los expertos se encuentran bastante concentrados y por tanto existe un gran acuerdo entre ellos. Por ejemplo, la Fig. 3 muestra una representación de las respuestas dadas por los meteorólogos para la dirección cardinal "oeste de Galicia" y un mapa de contornos que ilustra el porcentaje de respuestas que se solapan.

V. Creación de descriptores geográficos borrosos

En la Figura 3, la gráfica de contornos puede tomarse como la base de la semántica de su expresión correspondiente,

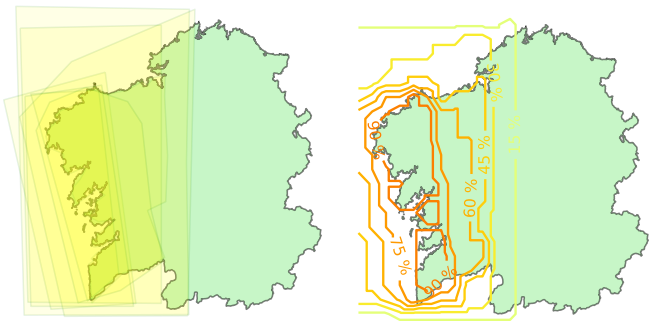

Fig. 3. Representación gráfica de los polígonos dibujados por los expertos y gráfica de contornos para "oeste de Galicia".

con una región nuclear que es aceptda por la mayoría, y un decaimiento gradual a medida que se avanza hacia la periferia exterior de las líneas de contorno. Así pues, en nuestro caso, la imprecisión surge de las diferencias interpersonales entre los meteorólogos.

Siguiendo esta noción, hemos creado modelos borrosos que agregan las opiniones de los expertos para cada descriptor. El método que usamos para esta tarea de modelado es una mejora del algoritmo heurístico descrito en [30], que producía modelos básicos basado en un muestreo de puntos y el conteo de intersecciones de polígonos, sin la inclusión de ninguna condición previa. Nuestro algoritmo va precedido además por un filtrado simple de los polígonos.

A. Filtrado de datos atípicos

Como comentamos anteriormente, los polígonos dibujados por los meteorólogos son muy consistentes visualmente, pero en algunos casos hemos observado pequeñas inconsistencias. Para mantener la elevada consistencia entre las respuestas obtenidas para los descriptores, aplicamos un filtrado simple de los trazados, consistente en descartar respuestas fuera del intervalo [media $\pm$ 2*desviación típica] en términos de tamaño y localización del centroide.

B. Caracterización de un descriptor geográfico borroso

Usando los datos filtrados, pretendemos construir descriptores geográficos borrosos que sean simples y consistentes.

Definition 1. Descriptor geográfico borroso, G:

$$G = \{S, K, \mu_G\} \tag{1}$$

Formalmente, definimos un descriptor geográfico borroso G (ej. "sur de Galicia") como un conjunto de 3 elementos: un área de soporte S, un área de núcleo K, y una función de pertenencia borrosa μ_G, que evalúa el grado en el que un punto en un mapa ($p = (x, y) \mid x, y \in \mathbb{R}$) puede considerarse parte de G:

$$\mu_G : \{\mathbb{R}, \mathbb{R}\} \rightarrow [0, 1] \tag{2}$$

Tomando μ_G[1] como base, K y S pueden definirse como:

$$K = \{p\} \mid \mu_G(p) = 1 \tag{3}$$

$$S = \{p\} \mid \mu_G(p) > 0 \tag{4}$$

Así pues, K es el conjunto de puntos (o región) cuyos grados de pertenencia son máximos con respecto a G, mientras que el soporte incluye a todos los puntos con un grado de pertenencia mayor que cero. Sin embargo, para la consecución de modelos consistentes, necesitamos aplicar las siguientes restricciones a G:

$$\forall\ G, K \subseteq S \tag{5}$$

$$\begin{gathered} \forall\ \{p_i, p_j\} \mid p_i, p_j \in S,\ p_i, p_j \notin K \text{ y } p_i \neq p_j \\ d(p_i, K) > d(p_j, K) \Longleftrightarrow \mu_G(p_i) < \mu_G(p_j) \end{gathered} \tag{6}$$

Estas condiciones aseguran que los modelos borrosos sean consistentes, al evitar la posibilidad de obtener Ks y Ss disjuntos, y asegurar la monotonicidad para μ_G, donde $d(p, K)$ es la distancia euclídea a K desde un punto p en S.

C. Construyendo descriptores sobre los datos de expertos

Un descriptor geográfico borroso G, como "norte de Galicia", se modela de acuerdo con los polígonos dibujados por los expertos para dicha expresión. Formalmente, la colección de polígonos dibujados para un G específico se define como:

$$R_G = \{P_1, P_2, ..., P_n\} \tag{7}$$

Cada P representa un polígono, y n es el número total de polígonos restantes tras el filtrado inicial. Cada polígono se compone de un conjunto de vértices, definidos bajo una proyección *plate carrée* (pares de valores de longitud y latitud).

La primera tarea para modelar un G dado consiste en determinar sus constituyentes (S y K), puesto que calcular ambos nos permitirá caracterizar μ_G posteriormente. Para ello, en primer lugar se transforman las coordenadas de los polígonos en R_G desde una proyección *plate carrée* a una proyección *Mercator*. A continuación, se define una malla de puntos equidistantes sobre ambos ejes cartesianos, $D = (p_1, ..., p_i, ..., p_{|D|})$, que se encuentra delimitada por la extensión máxima de la geografía subyacente (en nuestro caso, la región de Galicia). La distancia entre los puntos de la malla viene determinada por un parámetro δ, que especifica un porcentaje de la extensión total del mapa. Por ejemplo, $\delta = 1$ significa que la distancia entre un par de puntos de la malla contiguous es igual al 1% de la distancia entre límites opuestos del map sobre uno de los ejes.

Como se especifica en el Algoritmo 1, usando D y la colección de trazados expertos R_G, calculamos el número de veces que cada p en D está contenido en un polígono P en R_G, para determinar el porcentaje de polígonos que se solapan en un p dado. Basándonos en los porcentajes calculados para todos los puntos en D, determinamos K usando una aproximación de mayoría simple. Por tanto, PK está compuesto de

[1]Por simplicidad, nos referiremos a $\mu_G(p)$ en vez de $\mu_G(x, y)$.

Algorithm 1 Cálculo de K y S para un descriptor G

Entrada: D, R_G
Salida: K, S

$PCS \leftarrow ()$
for all $p_i \in D$ **do**
  $count \leftarrow 0$
  for all $P_j \in R_G$ **do**
    if $p_i \in P_j$ **then**
      $count \leftarrow count + 1$
    end if
  end for
  $pcs_i \leftarrow count/|R_G|$
  $PCS \leftarrow PCS \cup pcs_i$
end for
$PK \leftarrow \{p_i \in D \mid pcs_i > 0.5\}$
$PS \leftarrow \{p_i \in D \mid pcs_i > 0\}$
$K \leftarrow ConvexHull(PK)$
$S \leftarrow ConvexHull(PS)$
$bp \leftarrow \text{argmax}_{p_i \in S}(d(p_i, K))$
$op \leftarrow \text{argmin}_{p_i \in D \text{ y } p_i \notin S}(d(p_i, bp))$
$od \leftarrow d(K, op)$
return K, S, od

aquellos puntos cuyos porcentajes son $> 50\%$, y PS cubre todos los puntos donde el porcentaje es $> 0\%$. En vez de considerar S y K como colecciones de puntos en D (lo que PK y PS son), calculamos sus envolventes convexas, esto es, los polígonos convexos que delimitan sus áreas. Este proceso elimina cualquier área disjunta perteneciente a K que pueda aparecer por divergencias entre los polígonos dibujados, ya que el envolvente las agrega bajo una única área. Finalmente, calculamos od, la suma de la distancia del vértice de S más lejano a K (bp) y la distancia mínima de ese punto a otro externo a S (op), que será usada en la definición de μ_G, junto con K y S.

D. Evaluación de un descriptor geográfico borroso

Tomando como base los tres elementos devueltos por el Algoritmo 1, la función μ_G que evalúa un punto p se define como:

$$\mu_G(p) = \begin{cases} 1 & \text{si } p \in K \\ 1 - d(p, K)/od & \text{si } p \in S \text{ y } p \notin K \\ 0 & \text{si } p \notin S \end{cases} \quad (8)$$

La función de pertenencia μ_G se define siguiendo la condición de monotonicidad. Bajo dicha definición, todos los puntos en S tienen grados de pertenencia > 0 y este grado decrece a medida que el punto evaluado se aleja de K. En estas condiciones, la caracterización de un descriptor geográfico borroso G mediante la definición de su soporte y su núcleo nos permite crear modelos simples que son consistentes y fáciles de interpretar.

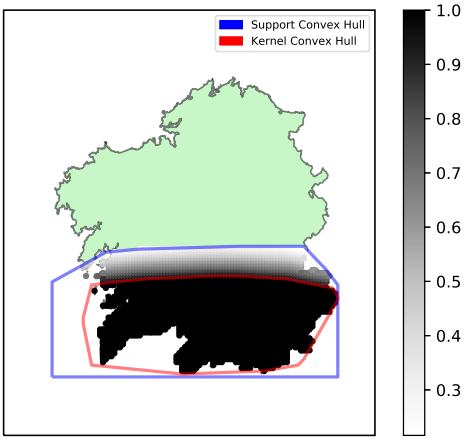

Fig. 4. Representación gráfica del descriptor geográfico borroso "Sur de Galicia".

E. Ejemplos

Con el fin de ilustrar los resultados del algoritmo de modelado y explicar algunas de las decisiones integradas en el Algoritmo 1 y en la definición de μ_G, describimos a continuación dos modelos borrosos distintos que resultan de la agregación de los polígonos dibujados por los meteorólogos para dos expresiones geográficas bajo distintas categorías o marcos de referencia en la Tabla I.

El primer descriptor, mostrado en la Figura 4, modela la expresión "sur de Galicia". Este modelo puede considerarse muy regular, tanto para K como para S. La condición de mayoría simple asegura además un mayor consenso a la hora de definir el significado de "sur". Este descriptor proporciona un área extensa en el que los puntos interiores pueden considerarse como parte del "sur de Galicia" en distintos grados.

Otros descriptores adoptan distintas formas, donde S varía en ancho respecto a K. Este es el caso, por ejemplo de "áreas de montaña de Lugo". Como muestra la Figura 5, la forma de este descriptor es en cierto modo elipsoidal y la distancia

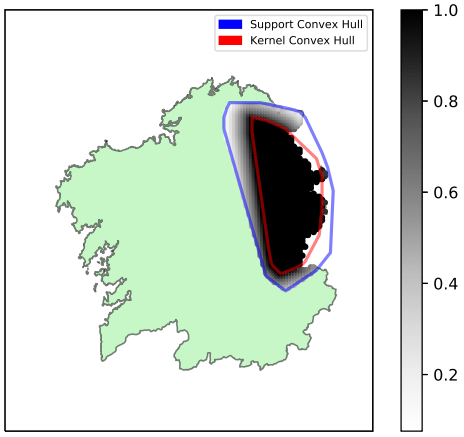

Fig. 5. Representación gráfica del descriptor geográfico borroso "Áreas montañosas de Lugo".

entre los límites de S y K no es constante, comparada con la Figura 4. Este caso ilustra bien cómo algunos puntos en S más cercanos a su periferia tienen grados de pertenencia más elevados que otros más lejanos, debido a su cercanía a K.

VI. Conclusiones

En este trabajo hemos descrito una metodología basada en la aplicación de conjuntos borrosos para la realización de una tarea de modelado de lenguaje, en el contexto del desarrollo de un sistema D2T que generará descripciones meteorológicas georreferenciadas. Esta metodología incluye la recogida de datos expertos sobre referencias geográficas y su agregación mediante la construcción de descriptores borrosos. Como trabajo futuro, desarrollaremos un algoritmo de generación de expresiones de referencia geográficas que utilice los descriptores descritos en este trabajo, que constituirá además el núcleo del sistema D2T.

Agradecimientos

A. Ramos-Soto es investigador postdoctoral financiado por la "Consellería de Cultura, Educación e Ordenación Universitaria" (481B 2017/030) y J.M. Alonso es Investigador Ramón y Cajal (RYC-2016-19802). Además, este trabajo está parcialmente financiado por los proyectos TIN2017-90773-REDT (iGLN), TIN2017-84796-C2-1-R (BIGBISC), TIN2014-56633-C3-1-R (BAI4SOW) y TIN2014-56633-C3-3-R (ABS4SOW) cofinanciados por el "Ministerio de Economía y Competitividad." También reconocemos el apoyo de la Xunta de Galicia (Centro singular de investigación de Galicia acreditació 2016-2019) y la Unión Europea (Fondo FEDER - European Regional Development Fund - ERDF).

Referencias

[1] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[2] A. Ramos-Soto, A. Bugarín, and S. Barro, "On the role of linguistic descriptions of data in the building of natural language generation systems," *Fuzzy Sets and Systems*, vol. 285, pp. 31–51, 2016.

[3] Gartner, "Neural Networks and Modern BI Platforms Will Evolve Data and Analytics," http://www.gartner.com/smarterwithgartner/nueral-networks-and-modern-bi-platforms-will-evolve-data-and-analytics/, accessed: 2017-03-14.

[4] S. Sripada, E. Reiter, and I. Davy, "Sumtime-mousam: Configurable marine weather forecast generator," *Expert Update*, vol. 6, no. 3, pp. 4–10, 2003.

[5] A. Ramos-Soto, A. Bugarín, S. Barro, and J. Taboada, "Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data," *Fuzzy Systems, IEEE Transactions on*, vol. 23, no. 1, pp. 44–57, Feb 2015.

[6] E. Goldberg, N. Driedger, and R. Kittredge, "Using natural-language processing to produce weather forecasts," *IEEE Expert*, vol. 9, no. 2, pp. 45–53, 1994.

[7] J. Coch, "Interactive generation and knowledge administration in multimeteo," in *Proceedings of the Ninth International Workshop on Natural Language Generation*, Niagara-on-the-lake, Ontario, Canada, 1998, pp. 300–303, software demonstration.

[8] J. Hunter, Y. Freer, A. Gatt, E. Reiter, S. Sripada, and C. Sykes, "Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse," *Artificial Intelligence in Medicine*, vol. 56, no. 3, pp. 157 – 172, 2012.

[9] E. Reiter, R. Robertson, and L. Osman, "Types of knowledge required to personalise smoking cessation letters," in *Artifical Intelligence and Medicine: Proceedings of AIMDM-1999*, W. Horn, Ed. Berlin, New York: Springer, 1999, pp. 398–399.

[10] A. Goldstein and Y. Shahar, "An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data," *Journal of biomedical informatics*, vol. 61, pp. 159–175, 2016.

[11] J. Yu, E. Reiter, J. Hunter, and S. Sripada, "Sumtime-turbine: A knowledge-based system to communicate gas turbine time-series data," in *Developments in Applied Artificial Intelligence*, ser. Lecture Notes in Computer Science, P. Chung, C. Hinde, and M. Ali, Eds. Springer Berlin Heidelberg, 2003, vol. 2718, pp. 379–384.

[12] N. Marín and D. Sánchez, "On generating linguistic descriptions of time series," *Fuzzy Sets and Systems*, vol. 285, pp. 6 – 30, 2016, special Issue on Linguistic Description of Time Series.

[13] R. de Oliveira, Y. Sripada, and E. Reiter, *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*. Association for Computational Linguistics, 2015, ch. Designing an Algorithm for Generating Named Spatial References, pp. 127–135.

[14] D. Roy and E. Reiter, "Connecting language to the world," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 1–12, 2005.

[15] E. Reiter, "An architecture for data-to-text systems," in *Proceedings of the 11th European Workshop on Natural Language Generation*, S. Busemann, Ed., 2007, pp. 97–104.

[16] J. Novikova, O. Dušek, and V. Rieser, "The E2E dataset: New challenges for end-to-end generation," in *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany, 2017, arXiv:1706.09254. [Online]. Available: https://arxiv.org/abs/1706.09254

[17] E. Reiter, S. Sripada, J. Hunter, and I. Davy, "Choosing words in computer-generated weather forecasts," *Artificial Intelligence*, vol. 167, pp. 137–169, 2005.

[18] E. Reiter and S. Sripada, "Should corpora texts be gold standards for nlg?" in *Proceedings of the International Natural Language Generation Conference*, 2002, pp. 97–104.

[19] R. Turner, S. Sripada, E. Reiter, and I. P. Davy, "Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data," *Applications and Innovations in Intelligent Systems*, vol. XV, pp. 75–88, 2007.

[20] J. Kacprzyk, "Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation," *IEEE Trans. Fuzzy Systems*, pp. 451–472, 2010.

[21] A. Ramos-Soto, A. Bugarín, and S. Barro, "Fuzzy sets across the natural language generation pipeline," *Progress in Artificial Intelligence*, pp. 1–16, 2016.

[22] R. Turner, S. Sripada, E. Reiter, and I. P. D. Davy, "Using spatial reference frames to generate grounded textual summaries of georeferenced data," in *Proceedings of the 2008 International Conference on Natural Language Generation (INLG08)*, Salt Fork, Ohio, 12-14 June 2008.

[23] S. C. Levinson, *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press, 2003, vol. 5.

[24] R. de Oliveira, S. Sripada, and E. Reiter, "Absolute and relative properties in geographic referring expressions," in *Proceedings of the 9th International Natural Language Generation conference*, 2016, pp. 256–264.

[25] P. Fisher, "Sorites paradox and vague geographies," *Fuzzy sets and systems*, vol. 113, no. 1, pp. 7–18, 2000.

[26] V. B. Robinson, "A perspective on the fundamentals of fuzzy sets and their use in geographic information systems," *Transactions in GIS*, vol. 7, no. 1, pp. 3–30, 2003.

[27] P. Fisher, A. Comber, and R. Wadsworth, "Approaches to uncertainty in spatial data," *Fundamentals of spatial data quality*, pp. 43–59, 2006.

[28] MeteoGalicia, "Meteogalicia's web site." http://www.meteogalicia.es, 2018.

[29] A. Ramos-Soto, N. Tintarev, R. de Oliveira, E. Reiter, and K. van Deemter, "Natural language generation and fuzzy sets: An exploratory study on geographical referring expression generation," in *IEEE World Congress on Computational Intelligence, 2016 IEEE International Conference on Fuzzy Systems,*, 2016.

[30] A. Ramos-Soto, J. M. Alonso, E. Reiter, K. van Deemter, and A. Gatt, "An empirical approach for modeling fuzzy geographical descriptors," in *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–6.