Tadeep: Traducción automática en profundidad

DATOS BÁSICOS DEL PROYECTO

Referencia:	**TIN2015-70214-P**
Modalidad: A / B	**B**
Investigador Principal 1:	**Kepa Sarasola Gabiola**
Organismo:	**Universidad del Pais Vasco (UPV/EHU)**
Centro:	**Facultad de Informática**

Subvención concedida (Costes directos): 96.800 €

Fecha inicio: 01/01/2016 Fecha finalización (prevista): 31/12/2018

Contrato predoctoral asociado: NO

Nombre: -----

Presenta: Kepa Sarasola (IP del proyecto)

PARTICIPANTES

Entidades participantes:

(1) Ixa Taldea, Fac. Informática, Universidad del Pais Vasco (UPV/EHU)
(2) Elhuyar Fundazioa
(3) Qatar Computing Research Institute

Equipo de investigación:

(1) Iñaki Alegria, Mikel Lersundi, Aingeru Mayor, Maite Oronoz, Kepa Sarasola (IP), Ruben Urizar.
+ Nora Aranberri (desde enero de 2017),
+ Izaskun Etxeberria (desde julio de 2016)
(2) Antton Gurrutxaga, Igor Leturia

Mujeres: 3 (30%)
Hombres: 7

Equipo de trabajo:

(1) Mikel Artetxe, Uxoa Iñurrieta, Gorka Labaka
(3) Lluis Màrquez

Mujeres: 1 (33%)
Hombres: 3

MOTIVACIÓN, HIPÓTESIS Y ESTRATEGIA DEL PROYECTO

Nos propusimos mejorar nuestros sistemas de Traducción Automática afrontando la integración de las redes neuronales y su aplicación por medio del "word embedding" y "deep-learning".

También queríamos adaptar las herramientas Depfix y TectoMT a la traducción entre lenguas en-es y en-eu (usando sintaxis profunda y semántica). Habíamos conocido esas herramientas en el proyecto europeo QT_LEAP.

TA adaptada a dominios específicos. Dadas las limitaciones de calidad en los sistemas de TA, una buena adaptación al dominio es una de las mejores garantías para la mejora de la calidad. Trabajaríamos en dominios técnicos como el dominio médico y el de consumo

Motivar científicamente el proyecto en el contexto de la línea de trabajo del equipo y de los resultados previos disponibles, describir la hipótesis de partida y la estrategia general del proyecto.

OBJETIVOS PROPUESTOS Y ALCANZADOS

1. **Traducción Automatica basada en análisis profundo**
   - Sistemas baseline: general y adaptado al dominio **(100%)**
   - Recursos adicionales para el dominio: corpora monolingües y bilingüe **(90%)**
   - Sistema de traducción de *OOVs* basado en *Word embedding* ***(100%)***
   - *Sistema de traducción basado en redes neuronales.* ***(100%)***
   - *Gramáticas TectoMT adaptadas a dominios, a partir de las ya disponibles en el proyecto QTLeap.* ***(100%)***
2. **Traducción Automatica adaptada a dominios específicos.**
   - Sistema general mejorado basado en morfologia y caracteres. **(100%)**
   - Sistema mejorado adaptado al dominio médico. **(80%)**
   - Sistema mejorado adaptado al dominio de consumo. **(100%)**

No necesitamos prórroga

RESULTADOS CIENTÍFICO-TÉCNICOS

1. Traducción Automatica basada en análisis profundo

Objetivo 1.3 Desarrollo de un sistema de traducción de *OOVs* basado **en** *Word embedding.*
Progreso: 100%
- Se ha desarrollado un sistema capaz de mapear embeddings monolingües en un mismo espacio vectorial, basándose únicamente en pequeños diccionarios bilingües de los pares de lengua involucrados.
- Los mejores resultados publicados en inferencia de diccionarios, incluso usando diccionarios mucho más reducidos a los habitualmente utilizados.
- Muy buenos resultados. Publicaciones:
    - A robust self-learning method for fully unsupervised cross-lingual mappings of word embedding. (ACL 2018)
    - VecMap: A framework to learn bilingual word embedding mappings (https://github.com/artetxem/vecmap)

Science journal: 'Ixa opens a new research avenue: Machine Translation without a dictionary?'

Atsegin dut 4 lagunek atsegin dute. Izena eman zurei lagunei gustazen zaiena ikusteko

Science reported this week about the work recently published by our colleagues Mikel Artetxe, Eneko Agirre and Gorka Labaka: **Artificial intelligence goes bilingual—without a dictionary**

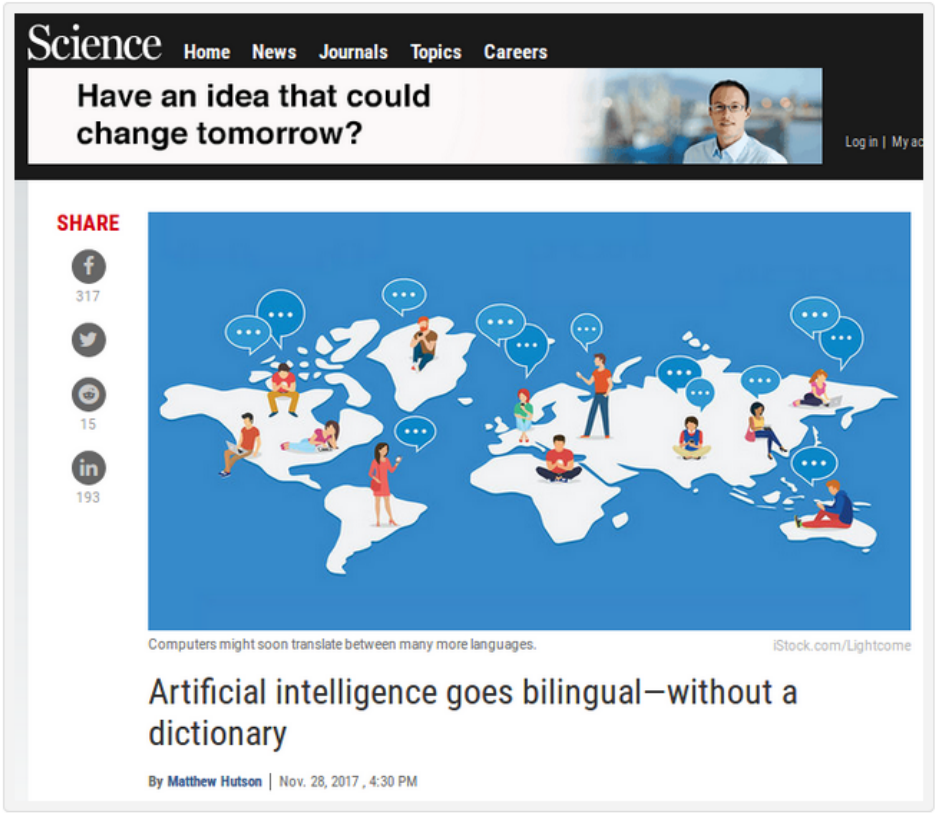
In October the 30th our three colleagues published a pre-print paper entitled **Unsupervised Neural Machine Translation** in collaboration with **Kyunghyun Cho**.

One day later G. Lample published another paper with similar contents entitled **Unsupervised Machine Translation Using Monolingual Corpora Only**. Both papers are under consideration at **ICLR 2018**.

Those are some sentences written by **Matthew Hutson** a freelance writer covering technology for Science:

> *[…] two new papers show that neural networks can learn to translate with no parallel texts—a surprising advance that could make documents in many languages more accessible.*
>
> *[…] Imagine that you give one person lots of Chinese books and lots of Arabic books—none of them overlapping—and the person has to learn to translate Chinese to Arabic. That seems impossible, right?" says the first author of one study, Mikel Artetxe, a computer scientist at the University of the Basque Country (UPV) in San Sebastián, Spain. "But we show that a computer can do that."*
>
> *[…] "This is in infancy," Artetxe's co-author Eneko Agirre cautions. "We just opened a new research avenue, so we don't know where it's heading."*
>
> *[…] Artetxe says the fact that his method and Lample's—uploaded to arXiv within a day of each other—are so similar is surprising. "But at the same time, it's great. It means the approach is really in the right direction."*

Congratulations Mikel, Eneko, Gorka and Kyunghyun!



http://www.sciencemag.org/news/2017/11/artificial-intelligence-goes-bilingual-without-dictionary

RESULTADOS CIENTÍFICO-TÉCNICOS

1. Traducción Automatica basada en análisis profundo

Objetivo 1.4 Desarrollo de un sistema de traducción basado en aprendizaje en profundidad (redes neuronales).

- Progreso: 100%
- Se ha generado una primera versión de los traductores neuronales (http://ixa2.si.ehu.es/nmtdemo) que se basan en la arquitectura propuesta por Bahdanau *et al.* (2015) y la segmentación automática de palabras basada en BPE (Sennrich *et al.*, 2016).
- Para tratar a morfología compleja del euskara, se ha creado otro sistema basado en caracteres.

Muy buenos resultados, recientemente publicados:

- *Neural Machine Translation of Basque.*
  European Association of Machine Translation (EAMT, 2018).

HOME | PARTICIPANTS | PUBLICATIONS | DEMOS

Demo

Ixa NMT: http://ixa.si.ehu.eus/nmtdemo/

[login]

Languages

- Euskara
- English
- Español

News

PROCESSING OF HISTORICAL CORPORA (Open day workshop, 2018-06-11)

Be a friend of the Minority SafePack! We need your signature!

Science journal: 'Ixa opens a new research avenue: Machine Translation without a dictionary?'

Course: Deep Learning for Natural Language Processing

http://ixa2.si.ehu.es/tadeep/demo

RESULTADOS CIENTÍFICO-TÉCNICOS

1. Traducción Automatica basada en análisis profundo

Objetivo 1.5 Desarrollo de Gramáticas TectoMT adaptadas a dominios, a partir de las creadas en el proyecto QTLeap.

- Progreso: 100%
- Incorporado al repositorio de TectoMT (http://github.com/ufal/treex).
- Generados los módulos para euskara y español.
- Modificados los módulos de inglés para poder traducir en cuatro direcciones (en->es, es->en, en->eu, eu->en)
- Buenos resultados, publicaciones:
    - Tectogrammar-based machine translation for English-Spanish and English-Basque (SEPLN, 2016)

El éxito de la NMT nos ha hecho postergar esta linea de momento.

RESULTADOS CIENTÍFICO-TÉCNICOS

Además:

Objetivo 1.2
Recopilación de recursos: corpora monolingües y bilingües, incluyendo la extracción de entidades nombradas, terminología... **(90%)**

Objetivo 2
Traducción Automatica adaptada a dominios específicos.

- Sistema mejorado adaptado al dominio médico. **(80%)**

  KabiTermICD: Nested Term Based Translation of the ICD-10-CM into a Minor Language. Multilingual Biomedical Text Processing 2018. Japan

- Sistema mejorado adaptado al dominio de consumo. **(100%)**

  Demo: *http://democonsumer.elhuyar.eus/*

RESUMEN DE LOS RESULTADOS DEL PROYECTO

	Número	Indicios de calidad
Artículos científicos derivados del proyecto en revistas JCR (de ellos cuantos en acceso abierto)	3 (0)	1 Q1; 2 Q4
Revisiones (surveys), editoriales y otros artículos científicos (de ellos cuantos en acceso abierto)	-	
Libros, capítulos de libros y monografías (nac/internac)	1	Capitulo de libro internacional
Conferencias en congresos (nacionales/internac, indicando cuántas por invitación).	1 / 15	GII-GRIN-SCIE (4 C1; 2 C2; 4 C3) CORE (3 A*; 3 A; 1 B; 4 C)
Patentes/Registros Software (indicar estado)	1 (trámite)	

Indique únicamente aquellos resultados que derivan directamente del presente proyecto

FORMACIÓN DE PERSONAL

Tesis doctorales realizadas relacionadas con el proyecto (con indicación de título, fecha de inicio y de lectura, e indicadores relativos a publicaciones derivadas)

Tesis: Izaskun Etxeberria, profesora del proyecto. 11-07-**2016**. Dir.: Iñaki Alegria Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz
Tesis: Olatz Perez de Viñaspre, tesis dirigida por Maite Oronoz (**2017**) Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea
Tesis a defender en 2018 o 2019:
- Uxoa Iñurrieta "Las colocaciones en la traducción automática"
- Mikel Artetxe. Neural Machine Translation sin información bilingüe

Otras tesis:
- Manex Agirrezabal (2017) Automatic Scansion of Poetry
- Xabier Saralegi (2017) Cross-Lingual Information Retrieval para lenguas con pocos recursos

Actividades de formación de predoctorales y/o personal técnico relacionadas con el proyecto (destacar / agregar)

Impartimos formación en el **Master Erasmus-Mundus** Language Analysis and Processing
Mikel Artetxe, becario FPI. Tesis de máster. Dir.: G. Labaka y E. Agirre 2016
Distributional Semantics and Machine Learning for Statistical Machine Translation
Estancias en New York con el prof. Cho (2017) y en Facebook (2018)
Usoa Iñurrieta. becaria FPI. Estancia en Inglaterra con John Carroll (2016)
Xabier Soto acaba de conseguir una beca para hacer una tesis en traducción automática para el dominio médico. (2018-2022)

INTERNACIONALIZACIÓN DE LA INVESTIGACIÓN

- En 2016 participamos en el proyecto europeo QTLeap (qtleap.eu) **QTLeap: Quality Translation by Deep Language Engineering Approaches**
  Financiación recibida (en euros): 385.270 (grupo) 3.003.540 (total)
- Organización del Open Workshop on Neural Machine Translation. with Kyunghyun Cho (2017-05-29)
- Estancia de 4 días en Donostia de Kyunghyun Cho (mayo de 2017)
- Estancia de Mikel Artetxe en 2017 en New York con Kyunghyun Cho,
- Estancia de Mikel Artetxe en 2018 en Facebook.
- Estancia de Usoa Iñurrieta en 2016 Sussex (John Carroll)
- Estancia de Kepa Sarasola en 2018 en Dublin (DCU, Andy Way)

Relacionar las colaboraciones internacionales relacionadas con el proyecto y la relevancia para el mismo y para el equipo investigador
Participación en proyectos europeos e internacionales, indicando Programa, convocatoria, tipo de proyecto y resultado de la propuesta presentada. Indicar financiación recibida en el grupo.

INTERNACIONALIZACIÓN DE LA INVESTIGACIÓN

- Aprobado proyecto europeo Interreg Poctefa (2018-2010):
  LINGUATEC: Desarrollo de la cooperación transfronteriza y la transferencia de conocimiento en tecnologías de la lengua.
  Presupuesto total: 797.875€.
  Consorcio: 1) Elhuyar Fundazioa; 2) Lo Congrès Permanent De La Lenga Occitana (Francia); 3) Universidad Del País Vasco / Euskal Herriko Unibertsitatea; 4) CNRS (CENTRE National De La Recherche Scientifique)- Delegation Regionale Midipyrenees (Francia); 5) Euskaltzaindia. Real Acaldemia De La Lengua Vasca; 6) Sociedad De Promoción Y Gestión Del Turismo Aragonés, S.L.U,

Relacionar las colaboraciones internacionales relacionadas con el proyecto y la relevancia para el mismo y para el equipo investigador
Participación en proyectos europeos e internacionales, indicando Programa, convocatoria, tipo de proyecto y resultado de la propuesta presentada. Indicar financiación recibida en el grupo.

Resultados y relaciones con el entorno socio económico

- Creación de un consorcio local para la explotación del sistema de traducción neuronal creado con éxito.
- Colaboración con un consorcio local para investigar sobre estimación automática de calidad de traducción:
  Proyecto QUALES: Aprendizaje Automático mediante Supervisión Modulable para Estimación Automática de Calidad de Traducción
- Colaboración y consultoría con el experto en Machine Learning Lluís Màrquez. Seminarios de seguimiento todos los años una semana en junio.
- Colaboración con otros proyectos:
  Europa: QT_LEAP
  Ministerio: PROSAMED, TUNER,
  País Vasco: MODELA, Eroski-Consumer

Relacionar (si procede) EPOs y su colaboración o papel en el proyecto. Relacionar colaboraciones, proyectos o contratos de I+D+i con el sector privado como resultado de la investigación. Relacionar participación (si procede) en iniciativas empresariales (p.e. spin-off). Indicar fuente de financiación y cuantía. Relacionar (si procede) patentes y el grado de explotación. Otras acciones de transferencia de tecnología (convenios, colaboraciones, consultorías, prestación de servicios, etc.)

Otros aspectos destacables relacionados con el proyecto

- **Artículo de la revista Science sobre la tesis de Mikel Artetxe en traducción sin comocimient blingüe**
- **Demo sistema NMT general**
  - Muy buenos resultados
  - Plan de explotación
- **Demo sistema NMT en dominio consumo**
  - ***http://democonsumer.elhuyar.eus/***
- **Open Workshop on Neural Machine Translation. with K yunghyun Cho (2017-05-29)**
- **Diseminación:** más de 20 noticias en los medios

P.ej. Organización de eventos, actividades de divulgación, etc.

EJECUCION DEL PRESUPUESTO

Concepto	Ejecutado: Cantidad y (%)	Existen cambios relevantes respecto a solicitud original? (*)
Inventariable	4.346 € (4,5%)	NO
Personal	64.842 € (67%)	NO
Fungible	0€ (0%)	NO
Viajes y dietas	3.047 € (3,1%)	NO

En esta diapositiva se indicará cómo se ha ejecutado (o está ejecutando) el gasto en relación con la solicitud presentada. No hace falta que las cantidades sean exactas, sino indicativas de la ejecución del gasto realizado hasta el momento

(*) Gastos no contemplados en la solicitud original

Sin modificaciones relevantes

PLANTEAMIENTO FUTURO

Seguir profundizando en estos aspectos:

- **Mejoras en traducción neuronal** (NMT):
  otras arquitecturas neuronales, morfología, adaptación al dominio, integración de morfologia. Dominio médico.

- **Problemas de NMT a resolver**. Falta de fidelidad: excesiva simplificación, fallos en entidades (números), y adjetivos (izda/dcha)...

- **Investigación en traducción sin información bilingüe**:
  Nuestro grupo está muy bien posicionado en esta importante tarea

- Adaptación al dominio utilizando datos libres de **Wikidata y Wikipedia**

Si tiene previsto solicitar proyecto en una próxima convocatoria, explique la posible relación (continuidad de la investigación) y diferencias con el presente proyecto.