CrowdTransEval: evaluación de sistemas de traducción automática mediante crowdsourcing

Alejandro Navarro Fulleda
Dirigido por: Juan Antonio Pérez Ortiz

Defensa de proyecto de final de carrera, 2012

Indice

Motivación del proyecto

La evaluación de traductores automáticos requería de:

- Expertos traductores → Evaluaciones costosas
  - Tiempo
  - Dinero
- Tareas casi idénticas → Necesidad de reutilización

Motivación del proyecto

La evaluación de traductores automáticos requería de:

- Expertos traductores $\rightarrow$ Evaluaciones costosas
  - Tiempo
  - Dinero
- Tareas casi idénticas $\rightarrow$ Necesidad de reutilización

Motivación del proyecto

La evaluación de traductores automáticos requería de:

- Expertos traductores $\rightarrow$ Evaluaciones costosas
  - Tiempo
  - Dinero
- Tareas casi idénticas $\rightarrow$ Necesidad de reutilización

Solución: Crowdsourcing

Definición

Consiste en externalizar tareas a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta.

- Tareas personalizadas
- Millones de posibles trabajadores
- Control de calidad
- Distintos servicios
  - Mechanical Turk (sólo para EEUU)
  - CrowdFlower (válido para Europa y la plataforma que usaremos)

Solución: Crowdsourcing

Definición

Consiste en externalizar tareas a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta.

- Tareas personalizadas
- Millones de posibles trabajadores
- Control de calidad
- Distintos servicios
  - Mechanical Turk (sólo para EEUU)
  - CrowdFlower (válido para Europa y la plataforma que usaremos)

Solución: Crowdsourcing

Definición

Consiste en externalizar tareas a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta.

- Tareas personalizadas
- Millones de posibles trabajadores
- Control de calidad
- Distintos servicios
  - Mechanical Turk (sólo para EEUU)
  - CrowdFlower (válido para Europa y la plataforma que usaremos)

Solución: Crowdsourcing

Definición

Consiste en externalizar tareas a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta.

- Tareas personalizadas
- Millones de posibles trabajadores
- Control de calidad
- Distintos servicios
  - Mechanical Turk (sólo para EEUU)
  - CrowdFlower (válido para Europa y la plataforma que usaremos)

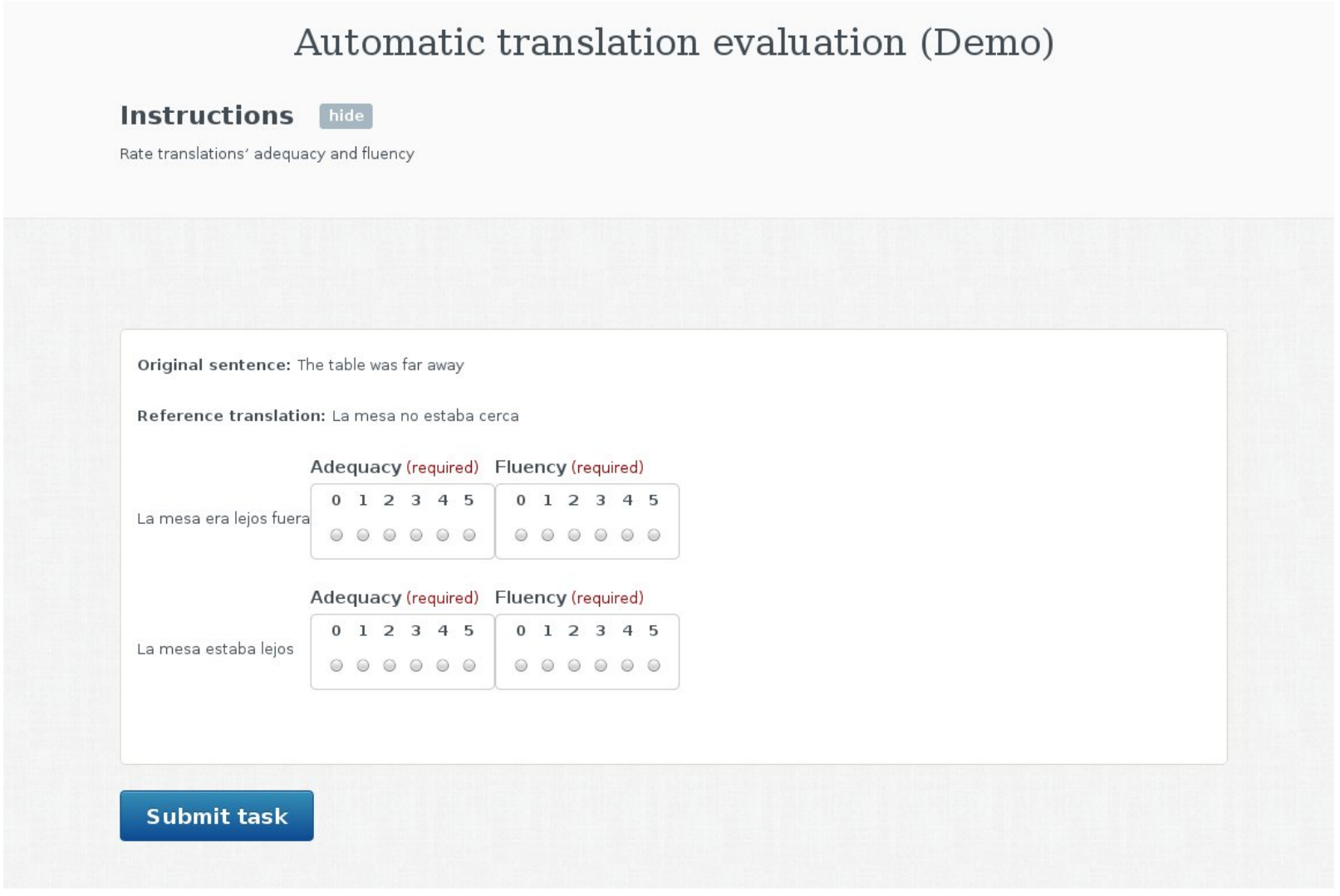
Solución: Crowdsourcing

Definición

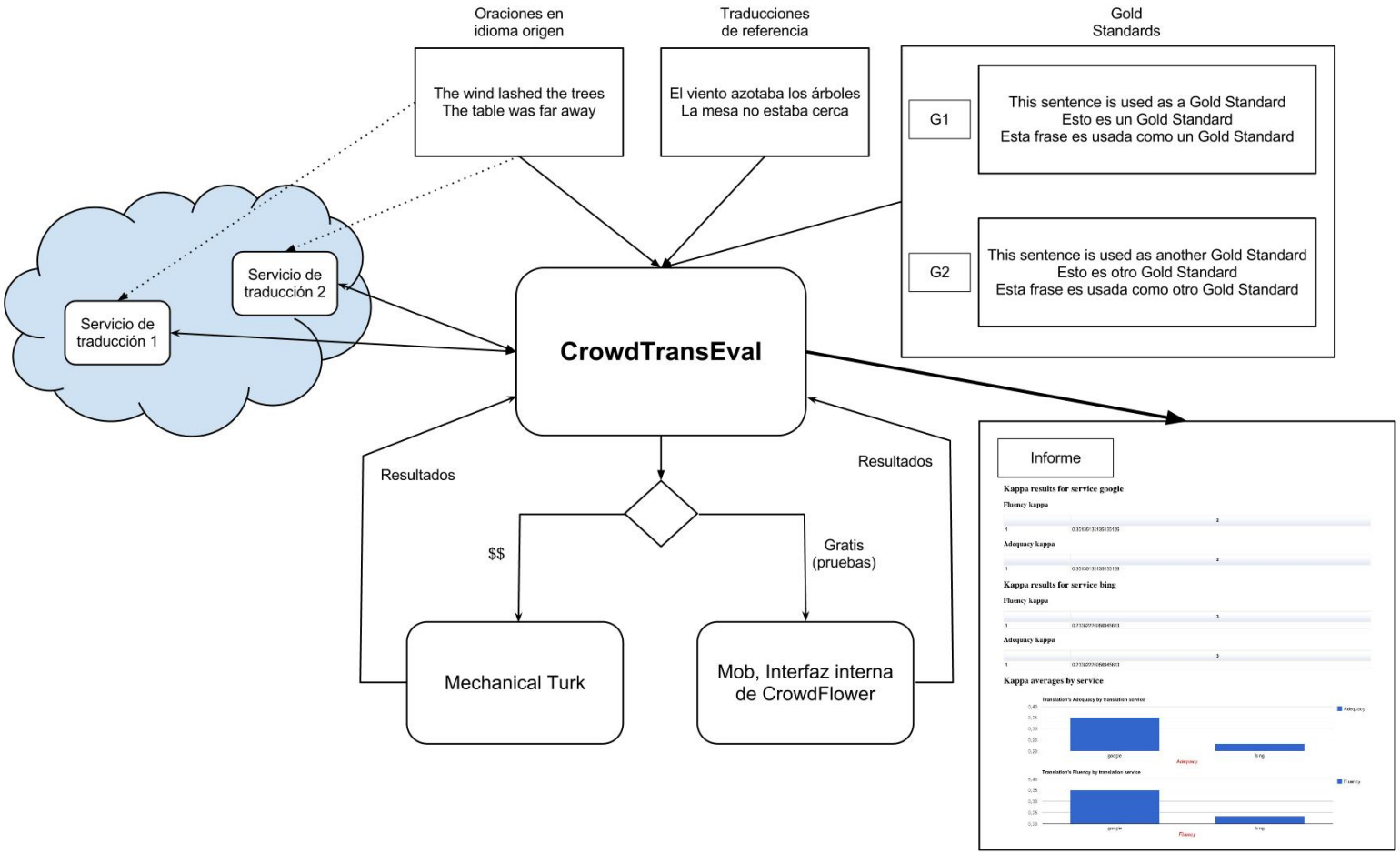
Consiste en externalizar tareas a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta.

- Tareas personalizadas
- Millones de posibles trabajadores
- Control de calidad
- Distintos servicios
  - Mechanical Turk (sólo para EEUU)
  - CrowdFlower (válido para Europa y la plataforma que usaremos)

Tarea ejemplo



Concepto de la aplicación



Indice

Jobs

- Son las tareas de evaluación
- Se componen de unidades a responder por el trabajador
- Se crean a partir de atributos definidos
  - Título
  - Descripción
  - Instrucciones

Jobs

- Son las tareas de evaluación
- Se componen de unidades a responder por el trabajador
- Se crean a partir de atributos definidos
  - Título
  - Descripción
  - Instrucciones

Jobs

- Son las tareas de evaluación
- Se componen de unidades a responder por el trabajador
- Se crean a partir de atributos definidos
  - Título
  - Descripción
  - Instrucciones

Units

- Representan una unidad de trabajo
  - En nuestro caso la evaluación de los distintos traductores para una frase
    - Fluidez de la traducción
    - Adecuación de la traducción

Units

- Representan una unidad de trabajo
  - En nuestro caso la evaluación de los distintos traductores para una frase
    - Fluidez de la traducción
    - Adecuación de la traducción

Units

- Representan una unidad de trabajo
  - En nuestro caso la evaluación de los distintos traductores para una frase
    - Fluidez de la traducción
    - Adecuación de la traducción

Gold standards

- Control de calidad ofrecido por CF
- Definen unidades con respuestas conocidas
- Mide la fiabilidad de un trabajador
  - Colocando trampas entre las unidades del trabajo
  - Desechando las respuestas de un trabajador poco fiable

Gold standards

- Control de calidad ofrecido por CF
- Definen unidades con respuestas conocidas
- Mide la fiabilidad de un trabajador
  - Colocando trampas entre las unidades del trabajo
  - Desechando las respuestas de un trabajador poco fiable

Gold standards

- Control de calidad ofrecido por CF
- Definen unidades con respuestas conocidas
- Mide la fiabilidad de un trabajador
  - Colocando trampas entre las unidades del trabajo
  - Desechando las respuestas de un trabajador poco fiable

Gold standards

- Control de calidad ofrecido por CF
- Definen unidades con respuestas conocidas
- Mide la fiabilidad de un trabajador
  - Colocando trampas entre las unidades del trabajo
  - Desechando las respuestas de un trabajador poco fiable

Indice

Indice

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Aplicación

- Aplicación por linea de órdenes
- Dos modos
  - Creación del trabajo
  - Demonio de comprobación de trabajo finalizado
- Podemos usar los dos en conjunto o por separado
- Necesita algunos datos para trabajar con las API's

Ambos modos

```
java -jar CrowdTransEval.jar
-cf config.properties -pf
parameters.properties -g
gold.txt -sl sl.txt -rt rt.txt -d
```

Sólo demonio

```
java -jar CrowdTransEval.jar -d
<id del trabajo a monitorizar>
```

Configuración de la aplicación

- Configuración a través de fichero de propiedades

Ejemplo de fichero config.properties

CrowdFlowerKey=<tu clave para el API de CF>
ShuffleGrade=1
Channels=mob
SL=en
TL=es
ApertiumKey=<tu clave para el API de Apertium>
BingClientId=<tu identificador de cliente para Bing translator>
BingClientSecret=<el secreto de la aplicacion de Bing translator>

Indice

Flujo del programa

1. Crear el trabajo en CF
2. Poblar el trabajo con unidades
3. Añadir *gold standards*
4. Encargar el trabajo
5. Esperar a que el trabajo termine
6. Recoger los resultados e interpretarlos

Creación del trabajo

- Configuración del trabajo mediante fichero de propiedades
- El trabajo se crea vacío para rellenarse posteriormente
- El identificador del trabajo se usará a través de todo el proceso

Creación del trabajo

- Configuración del trabajo mediante fichero de propiedades
- El trabajo se crea vacío para rellenarse posteriormente
- El identificador del trabajo se usará a través de todo el proceso

Creación del trabajo

- Configuración del trabajo mediante fichero de propiedades
- El trabajo se crea vacío para rellenarse posteriormente
- El identificador del trabajo se usará a través de todo el proceso

Ejemplo de configuración del trabajo

Ejemplo de fichero parameters.properties

```
title=Nombre del trabajo
instructions=Instrucciones de resolución del trabajo
judgment_per_unit=1
units_per_page=1
```

Rellenado del trabajo

- Dos ficheros con frases
  - sl.txt: las frases a traducir
  - rt.txt: traducciones de referencia de las frases
- Traducción real mediante los servicios de traducción
  - Uso de la API web de los mismos
- JSON para el transporte de datos

Rellenado del trabajo

- Dos ficheros con frases
  - sl.txt: las frases a traducir
  - rt.txt: traducciones de referencia de las frases
- Traducción real mediante los servicios de traducción
  - Uso de la API web de los mismos
- JSON para el transporte de datos

Rellenado del trabajo

- Dos ficheros con frases
  - sl.txt: las frases a traducir
  - rt.txt: traducciones de referencia de las frases
- Traducción real mediante los servicios de traducción
  - Uso de la API web de los mismos
- JSON para el transporte de datos

Rellenado del trabajo

- Dos ficheros con frases
  - sl.txt: las frases a traducir
  - rt.txt: traducciones de referencia de las frases
- Traducción real mediante los servicios de traducción
  - Uso de la API web de los mismos
- JSON para el transporte de datos

Rellenado del trabajo

- Dos ficheros con frases
  - sl.txt: las frases a traducir
  - rt.txt: traducciones de referencia de las frases
- Traducción real mediante los servicios de traducción
  - Uso de la API web de los mismos
- JSON para el transporte de datos

Ejemplo de ficheros de frases

Ejemplo de fichero sl.txt

The wind lashed the trees
The table was far away

Ejemplo de fichero tr.txt

El viento azotaba los árboles
La mesa no estaba cerca

Creación de Gold Standard

- Fichero gold.txt
  - Frase en idioma original
  - Traducción de referencia
  - Traducción real
- Creación de frases falsas
- JSON para subirlos al trabajo

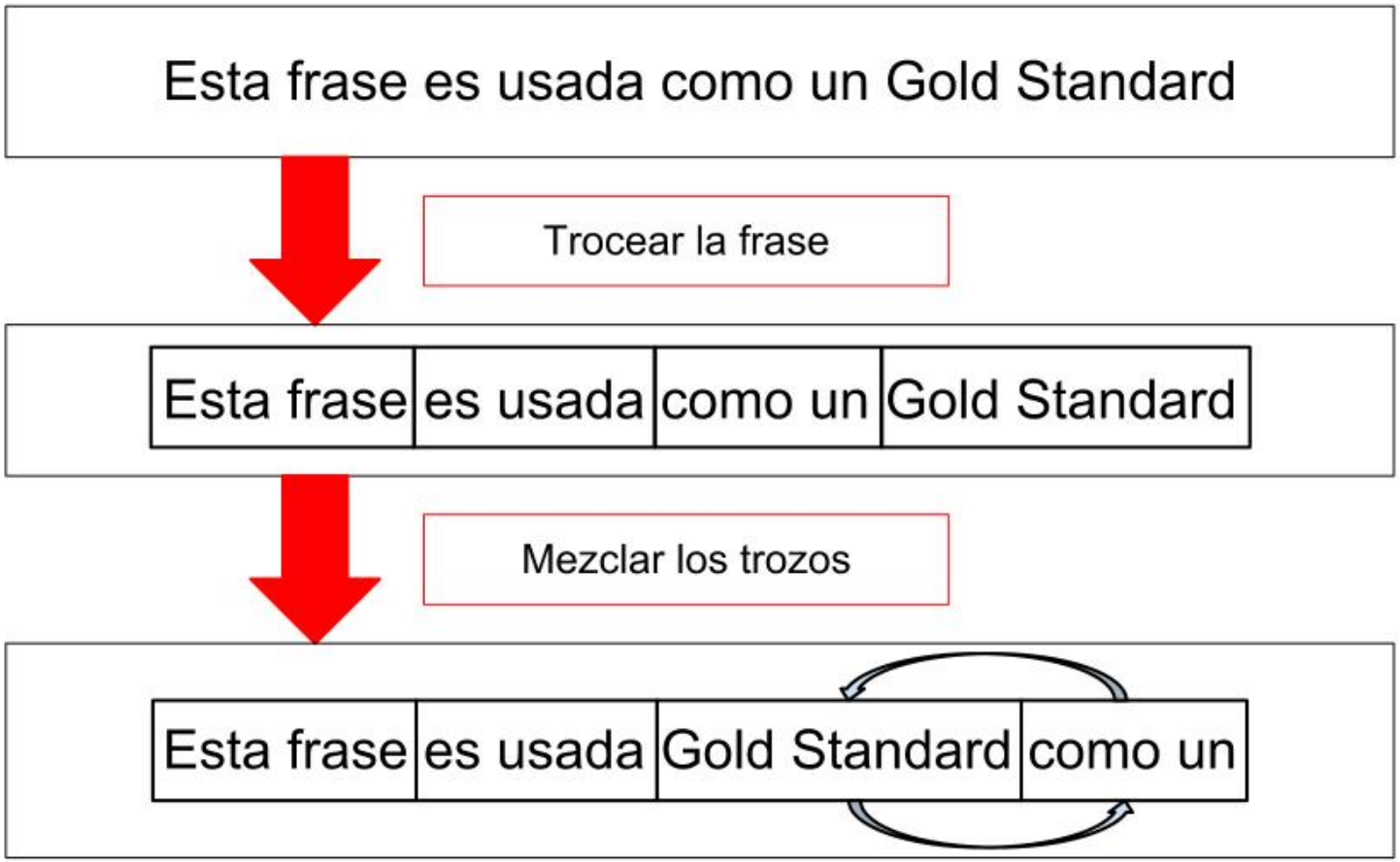
Creación de Gold Standard

- Fichero gold.txt
  - Frase en idioma original
  - Traducción de referencia
  - Traducción real
- Creación de frases falsas
- JSON para subirlos al trabajo

Creación de Gold Standard

- Fichero gold.txt
  - Frase en idioma original
  - Traducción de referencia
  - Traducción real
- Creación de frases falsas
- JSON para subirlos al trabajo

Barajado de frases

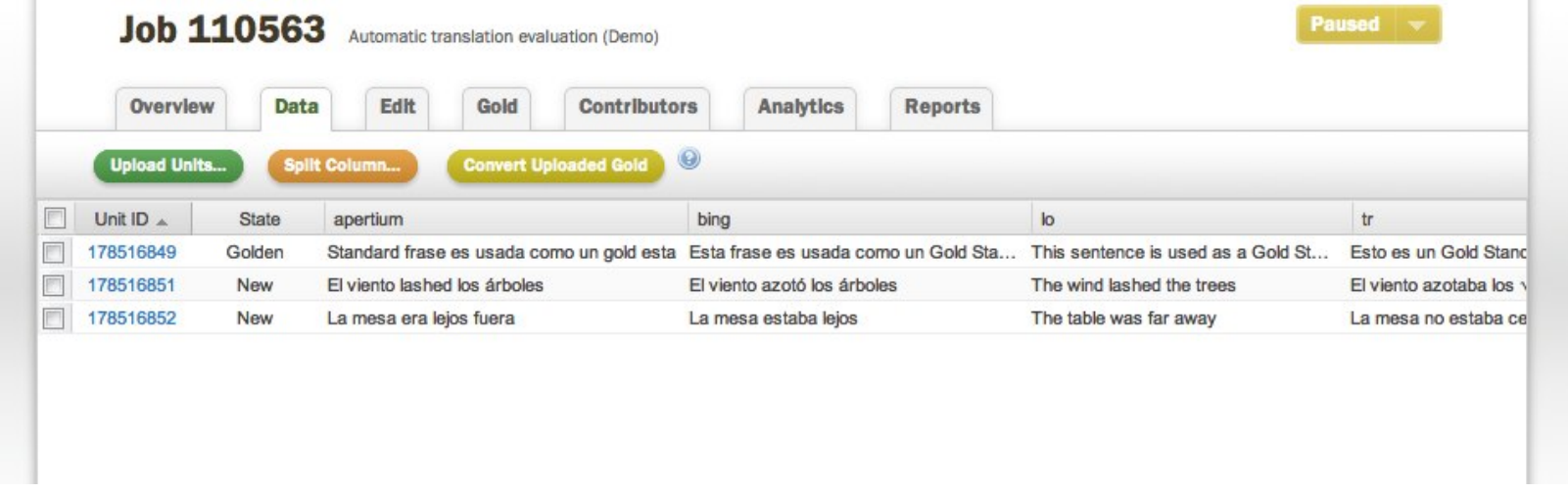


Ejemplo de fichero de Gold Standards

Ejemplo de fichero gold.txt

This sentence is used as a Gold Standard
Esto es un Gold Standard
Esta frase es usada como un Gold Standard

Ejemplo de trabajo con unidades



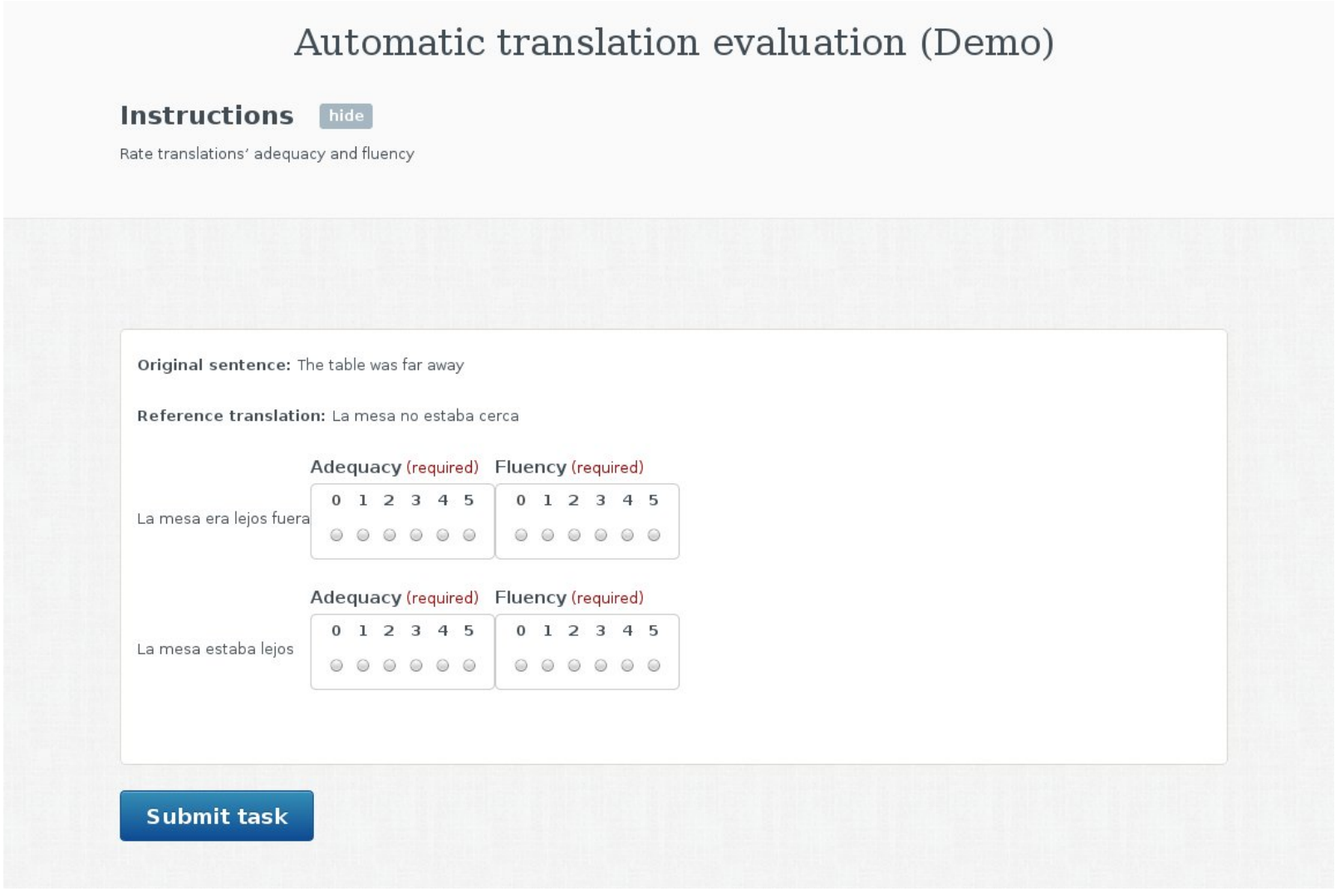
Encargar el trabajo

- Publicación del trabajo en los canales configurados
  - Propiedad Channels de fichero de configuración
- Mob, interfaz interna de CrowdFlower
  - Gratuita
  - Utilizada con el propósitos de realizar pruebas
  - Podemos resolver nuestros propios trabajos

Encargar el trabajo

- Publicación del trabajo en los canales configurados
  - Propiedad Channels de fichero de configuración
- Mob, interfaz interna de CrowdFlower
  - Gratuita
  - Utilizada con el propósitos de realizar pruebas
  - Podemos resolver nuestros propios trabajos

Ejemplo de resolución de trabajo



Indice

Generación de resultados

- Cuando se cumplen todos los juicios encargados
- Obtención de medidas
  - Puntuaciones medias de cada servicio
  - Valores de acuerdo entre anotadores (Factor Kappa de Cohen)

Generación de resultados

- Cuando se cumplen todos los juicios encargados
- Obtención de medidas
  - Puntuaciones medias de cada servicio
  - Valores de acuerdo entre anotadores (Factor Kappa de Cohen)

Factor Cohen's Kappa

Determina cuanto acuerdo existe entre pares de anotadores (trabajadores)

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Siendo:

- Pr(a): el acuerdo relativo entre pares de anotadores
- Pr(e): la probabilidad de acuerdo fortuito

Presentación resultados

- HTML con las medidas generadas
  - Google Chart Tools
- CSV con los resultados por unidad y trabajador

Ejemplo de informe de resultados

Kappa results for service google

Fluency kappa

	2
1	0.35135135135135126

Adequacy kappa

	2
1	0.35135135135135126

Kappa results for service bing

Fluency kappa

	3
1	0.23382226056945643

Adequacy kappa

	3
1	0.23382226056945643

Kappa averages by service