Language Switching A/B Test Results

Oliver Keyes

November 24, 2015

Summary: this was an A/B test to determine the impact of switching search languages in the case that a query produces zero results. Our hope was that applying language detection methods to failed queries would let us, for a subset, identify that they were using the wrong language, switch them over to the correct one, and get them results.

In practice, we found no evidence that this had a substantial impact on the zero results rate. Our hypothesis is that this is due to a combination of poor language detection and the small scale of queries that produced zero results and were not in the language of the project they were made on.

Background

One of the core goals of the Discovery team is to improve the Zero Results Rate - the proportion of user searches that return no results. After examining failed queries we noticed that some queries were perfectly valid, but simply in the wrong language; accordingly, we concluded that if we *detected* the user's language and re-ran their query, over indexes in that language, we would be able to reduce the zero results rate.

We launched a test that collected data from 4 November 2015 to 11 November 2015, running over 1 in 10 queries. Half of the search queries (the control group) were given the *status quo*; half (the test group) would have language detection methods and a second search query if their query produced 0 results. Over a week of testing this produced 25,348,283 events.

Exploratory analysis

Our first task is to explore the data to make sure, amongst other things, that sampling worked correctly. If it did not, the test group may not be comparable with the non-test group.

First, we looked at things on a by-day basis; for each day in the test, how many events do we have in each group?

Events by day and group, Language switching test
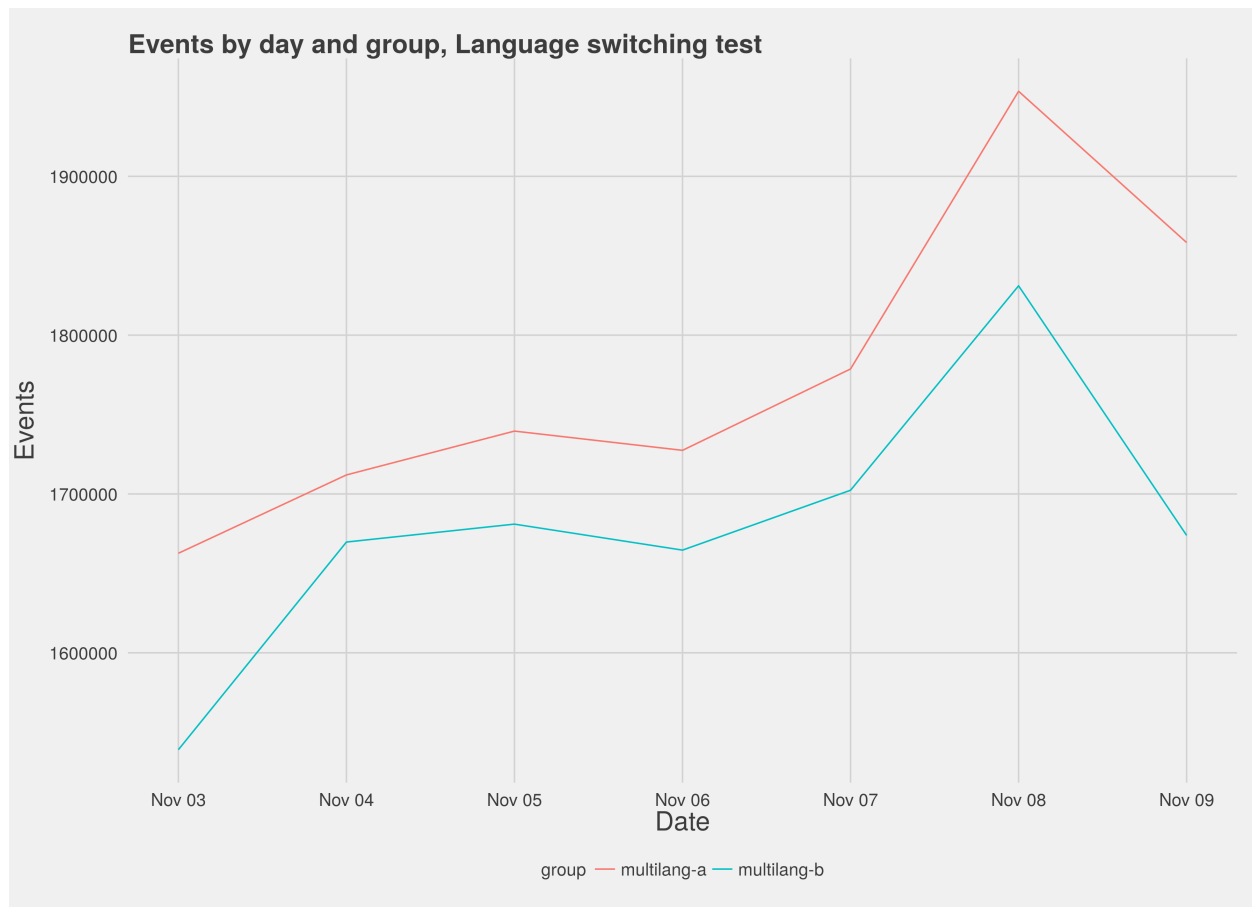
What this shows is that sampling is indeed uneven; the control group gets significantly more events than the test group (although both are still large enough to rely on).

Distribution of events per group and source, Language switching test
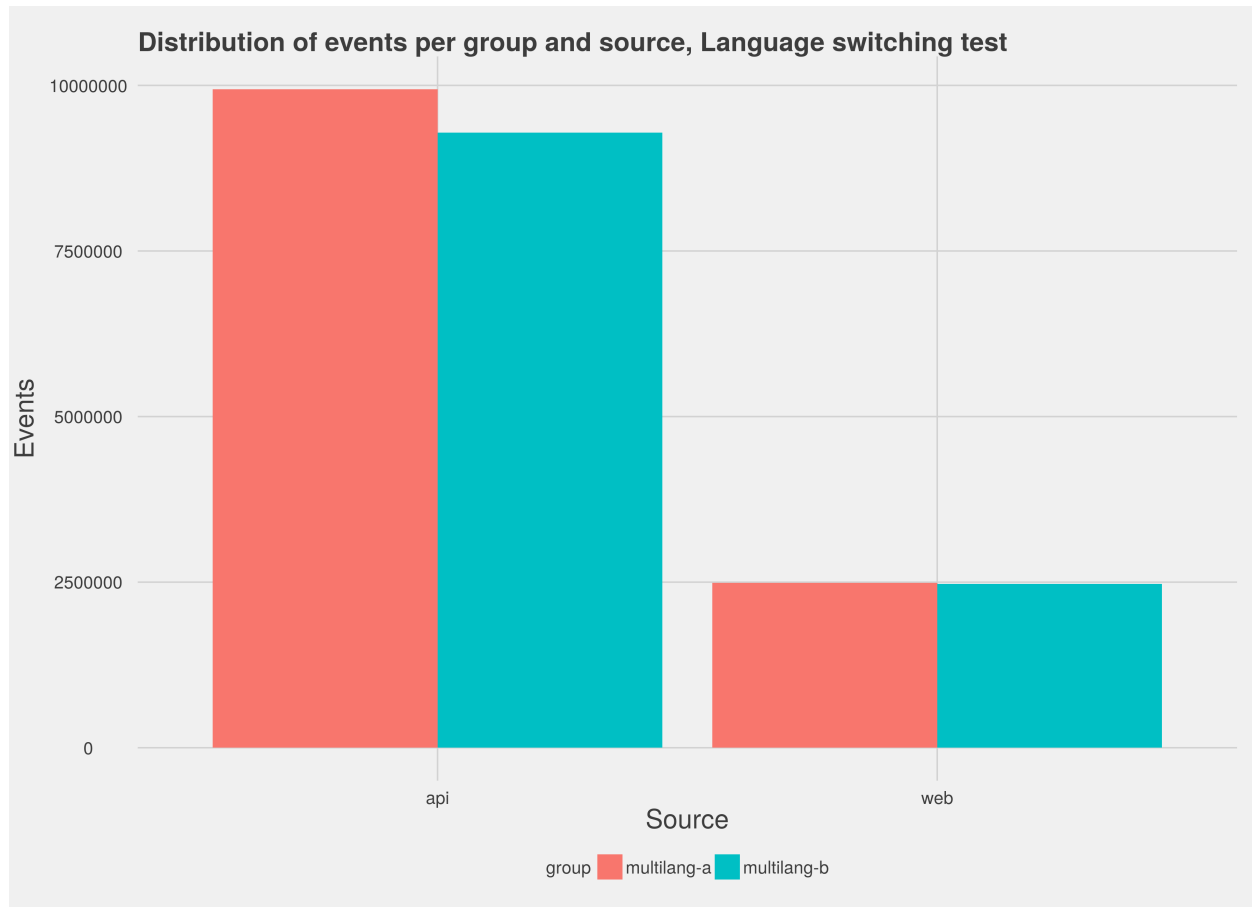
Diving deeper into the data we see that the discrepancy is caused by queries from the API; queries through the web interface (which are more likely to be "user" queries) are pretty much even. Accordingly we will look at both the outcome for the test as a whole, and the outcome for just "web" queries.

Results

To analyse the results we relied on Bayesian methods, since these don't depend on sample size restrictions and, absent those restrictions, are far less likely to produce an artificial and false result. We look at the "credible interval" for the difference in proportions between the test and control groups; in other words, whether the difference is big enough to be the result of a positive impact of the test rather than just statistical noise.

What we would like to see is a range that is:

1. Above 0, indicating that the outcome was positive;
2. Far enough 0 to indicate that the outcome was substantial.

With the overall data, we see a credible interval between **-0.5587085** and **-0.5580589**. This is not above 0, indicating that the outcome was not positive; the language test does not tend to result in an improved outcome for users.

If we look just at the web data, we see a credible interval between **-0.0004178118** and **0.0009659428**. *Some of this range is above 0, but not all of it, and the actual value above 0 is tiny.* Again, we cannot say there was a positive outcome for users as a result of this test.

Conclusion

The test did not show that language switching in the case that 0 results are returned produces a positive outcome for users. This may simply be because our language detection methods are naive - subjectively, we have seen them return unexpected and incorrect answers. It may also be that the failed queries are still not sensical even translated.

We recommend disabling the test and not implementing this feature as the default. Future work should look at the quality of language detection, or the actual queries that are being sent through language switching, in a qualitative way.